PhD THESIS

Models in molecular evolution: the case of toyLIFE

Pablo Catalán Fernández

UNIVERSIDAD CARLOS III DE MADRID

PhD THESIS

Models in molecular evolution: the case of toyLIFE

Author

Pablo Catalán Fernández

Supervisor

José Antonio Cuesta Ruiz

DEPARTMENT OF MATHEMATICS

Leganés, March 2017

Universidad
Carlos III de Madrid

Tesis doctoral

Models in molecular evolution: the case of toyLIFE

Autor
Pablo Catalán Fernández

Supervisor
José Antonio Cuesta Ruiz

Firma del Tribunal Calificador:

Presidente: Joshua L. Payne

Vocal: Jacobo Aguirre Araujo

Secretario: Saúl Ares García

Calificación:

Leganés, March 2017

a mis padres

Agradecimientos

Parece mentira. Llevo soñando con escribir estos agradecimientos desde que era adolescente. Sí, soy un raruno, lo sé. Mi madre y dos de mis tíos son doctores, y era evidente para mí que algún día yo también lo sería. Así, entre sueños con amores platónicos o mi éxito como guitarrista de un grupo de rock, me dedicaba a planear mi futura tesis doctoral. Y ninguna tesis está completa sin sus agradecimientos, claro. Al fin, la tesis de verdad llegó y acabó, y toca dar, de una vez, gracias.

Es de rigor —o al menos en mi cabeza lo es— empezar dando las gracias al director de tesis. De rigor o no, lo cierto es que mi gratitud hacia Jose Cuesta es grande y sincera, así que tenía que ser él el primero en aparecer. A lo largo de estos casi cinco años que nos conocemos, a Jose le ha dado mucho tiempo a reírse de (y con) mis ocurrencias, a maravillarme con su magia en la pizarra, y a, en definitiva, ser el mejor director de tesis que hubiese podido imaginar. Mi tesis no será brillante, y posiblemente no será el mejor trabajo que puedo hacer. Pero estos cinco años compartidos con Jose han estado llenos de aprendizaje, descubrimiento y ciencia. La tesis se acaba, pero la colaboración no. Así que, mal que le pese, seguiré dándole la tabarra mientras pueda. Gracias, Jose.

Esta tesis no habría sido posible sin la inestimable ayuda de Tito (también conocido como Clemente) Fernández. Literalmente no habría sido posible, pues él fue el artífice de toyLIFE, que constituye gran parte de este ladrillo que tenéis en las manos. De Tito he aprendido muchas cosas (buenas y malas), y soy realmente afortunado de haber podido verle en acción: tendré suerte si me vuelvo a encontrar a alguien tan brillante y apasionado. Ah, y antes de que se me olvide: el diseño gráfico de toyLIFE

también es su obra, así que si os gustan las figuras de la tesis ya sabéis a quién dar las gracias.

Siguiendo la ronda, toca agradecer a Susanna Manrubia, líder de nuestro grupo hermano en el Centro Nacional de Biotecnología. A medida que la tesis ha ido avanzando, la presencia de Susanna se ha hecho cada vez más patente, así como sus consejos y su cariño. Gracias, Susanna, ha sido (y seguirá siendo) un placer trabajar contigo.

Una tesis de este tipo requiere muchas horas en el despacho, haciendo simulaciones y estudiando matemáticas. No puedo creer la suerte que he tenido de haber compartido gran parte de estos años con los dos Ignacios (Pascual y Tamarit). Con vosotros he compartido ciencia, amistad, tontunas y muchas risas. Gracias por hacer de esta tesis una experiencia tan amena y agradable, chicos. ¡Habrá que quedar para tomar cañas de vez en cuando!

Además de mis compañeros de despacho, en estos años de tesis he pasado mucho tiempo con mucha gente que me ha tratado muy bien. He de dar las gracias a la gente de Susanna, destacando la presencia de Jacobo Aguirre, grande entre los grandes, pero también de Capi, Carlos, Adriana, Pilar y Toño. También siento profunda gratitud hacia nuestros vecinos de despacho (Misael, Ernesto, Juan Carlos y David), con quienes tantas comidas he compartido, y María Pereda, que en tan solo un año se ha convertido en una buena amiga con un corazón enorme. También tengo que dar gracias al resto de miembros del Grupo Interdisciplinar de Sistemas Complejos. He de hacer especial mención a Anxo Sánchez y Saúl Ares, con quienes he compartido muchos cafés matutinos y vespertinos, pero también a todos los demás miembros que se han portado tan bien conmigo: Rodolfo Cuerno, que me caía bien hasta que empecé a ir a aikido con él (y ahora me cae mejor), Yuri Martínez-Ratón, Carlos Rascón, Javi Muñoz, Esteban Moro, Silvia Santalla, Alberto Antonioni, y la gente de fuera de la UC3M como Luis Dinis, Juanma Parrondo, Ricardo Brito, Mario Castro, Ester Lázaro o Daniele Vilone. Y que no se me olviden Antonio García, Froilán Martínez, Alex Ortega y Alberto Calvo, del Departamento de Matemáticas. Gracias a todos por ser tan majos conmigo.

I would also like to thank Andreas Wagner for his hospitality and kindness and the members of his laboratory in Zurich, with whom I spent four weird months: Josh, Sinisa, José, Magda, Ali, Charles, Kathleen, Yolanda,

Fahad, Kasia, Debbie, Macarena and the rest. If I remember those months with a smile it is because of you guys. Thanks.

Given that I'm already writing in English, I might as well thank the people at Rob Beardmore's lab, including Rob himself, but also Carlos, Ivana, Cyrielle, Bogna, Peter, Lisa, Rick, Sarah and Emily. Those were an interesting couple of months. Thanks for that.

Antes de querer dedicarme a esto de la biología teórica (¿matemática? ¿cuantitativa?), estudié cinco años en la facultad de Biología en la Complutense. Aunque conocí a mucha gente allí, y agradecer a todo el mundo me llevaría hojas y hojas (básicamente, si hablaste conmigo durante esos cinco años, te doy las gracias), no puedo dejar de mencionar a Juan Antonio Delgado y a Mariló Jiménez, que fueron como mis papis en la investigación (aunque ya había hecho mis pinitos en el departamento de Genética de la mano de César Benito: gracias, Mario). Juan y Mariló, no solo fuisteis buenos jefes, sino mejores amigos, y os guardo un cariño inmenso que os demuestro menos de lo que debería. Os debo unas cuantas visitas y un abrazo. Esto tampoco sería posible sin vosotros. Gracias también a toda la gente del grupo de Luis Balaguer (que en paz descanse) y del Departamento de Ecología: Irene Cordero (esas sardinas quedarán en mi mente hasta el fin de los tiempos), Esther Pérez-Corona, Paloma de las Heras, Curro, Peri, Ana, Rocío, Sandra, Adri, Agus y los demás.

También le debo unas gracias a la gente del CCMA en el CSIC: MJ, Lilia, Raúl, Vanessa, Ana, Sergio, Irene otra vez...estuve poco tiempo por allí, pero me cuidasteis bien. Gracias, de verdad.

A todos mis profesores, desde el colegio hasta el doctorado, gracias. Supongo que de todos aprendí algo, aunque fuese poquito.

Ya sin relación directa con el trabajo, gracias a todos mis amigos. A los de toda la vida, como Guille y Luis y Dani, que siempre estarán ahí pase lo que pase y que son como hermanos para mi (si conseguís entender algo de la tesis os invito a cenar al VIPS). A los Feeder, que llegaron más tarde pero ya no se irán nunca: Carlos, Rudi, Adri y Gon. Lo que hemos vivido juntos, chicos. Y lo que nos queda. A los amigos con los que he pasado más tiempo este último año, como Paloma y Jose Ramón, o Eva y Raquel, o Diego: gracias por estar ahí. A la gente de biología, aunque apenas os veo ya: Javi, Lucía, las Martas, Esti, Marian, Sara, Tole... Y también a Aurora, Dani, Ceci, Miki y toda la gente de bota de la facultad.

A los Lamprologus. A mis maestros de kárate, Javier y Luiki, de quienes he aprendido tanto, y a mis compañeros: a todos los del Noru y los del San José del Parque, muchas gracias. A la gente de Zurich, por los momentos entrañables: Ugaitz, Andrea, Clara, Andrés, Nathalie... (Magda, a ti ya te he nombrado, no seas avariciosa!).

A David, mi hermano mayor perdido en las llanuras danesas. Que sigamos sentándonos juntos muchos años más. A la gente del Bosque Theravada, por todo lo que aprendí con ellos, y también a Jerôme Lamarlere y su gente. Thanks to Jeff and all his people: Bart and Fiona, Stefan and Guus, Ron, Karin, Ludwig and the rest. Special thanks to Fusako, who has been so kind, and to Haochen, my second lost brother, from China all the way to Switzerland! Big hugs to all of you. Thanks as well to Miles and all his family in Washington, who were so kind to me.

A mi familia, en especial a mis tíos Leonardo (y Raúl) y José Luis, que han seguido con extrañeza la tesis de su sobrino.

A mi hermano, con quien comparto esa complicidad que solo se consigue cuando compartes toda tu infancia con alguien. Gracias por escucharme y por inspirarme con tu ejemplo. Y por aguantarme todos estos años.

A mi padre, que me transmitió su pasión por aprender y que siempre me animó a dar lo mejor de mí mismo. A mi madre, que me animó a descubrir lo que me gustaba y a perseguirlo, y que nunca ha dejado de cuidarme. A los dos, porque os habéis asegurado de que nunca me faltase de nada, y porque me hicisteis saber que, pase lo que pase, alguien me va a seguir queriendo. Supongo que estaréis orgullosos de mí. Estad orgullosos de vosotros: si estoy aquí es porque vosotros lo hicisteis posible. Gracias, progenitores.

A Lucía, claro. He compartido contigo casi toda mi vida adulta, y aún sigo aprendiendo de tí. Gracias por todo el apoyo que me has brindado estos años, y que sigues brindándome justo ahora, escribiéndome mensajes de apoyo mientras redacto estas líneas. No ha sido fácil, pero ya se ha terminado. Te escribiría un mundo, pero me fallan las palabras. Gracias.

En definitiva, gracias a cualquier que me haya oído despotricar de la tesis y el desastre que es el mundo científico y la vida académica. Si tu nombre no está ahí y lees esto, dímelo y te compro una Coca-Cola (he dicho Coca-Cola y no cerveza, ¿vale? Que nos conocemos...).

En serio. Gracias.

Summary

This thesis set out to contribute to the growing body of knowledge pertaining models of the genotype-phenotype map. In the process, we proposed and studied a new computational model, tOYLIFE, and a new metaphor for molecular evolution —adaptive multiscapes. We also studied functional promiscuity and the evolutionary dynamics of shifting environments.

The first result of this thesis was the definition of tOYLIFE, a simplified model of cellular biology that incorporated toy versions of genes, proteins and regulation as well as metabolic laws. Molecules in tOYLIFE interact between each other following the laws of the HP protein folding model, which endows tOYLIFE with a simplified chemistry. From these laws, we saw how something reminiscent of cell-like behavior emerged, with complex regulatory and metabolic networks that grew in complexity as the genome increased.

tOYLIFE is, to our knowledge, the first multi-level model of the genotype-phenotype map, compared to previous models studied in the literature, such as RNA, proteins, gene regulatory networks (GRNs) or metabolic networks. All of these models either disregarded cellular context when assigning phenotype and function (RNA and proteins) or omitted genome dynamics, by defining their genotypes from high-level abstractions (GRNs and metabolic networks). tOYLIFE shares the same features exhibited by all genotype-phenotype maps studied so far. There is strong degeneracy in the map, with many genotypes mapping into the same phenotype. This degeneracy translates into the existence of neutral networks, that span genotype space as soon as the genotype contains more than two genes. There is also a strong asymmetry in the size distribution of phenotypes: most pheno-

types were rare, while a few of them covered most genotypes. Moreover, most common phenotypes are easily accessed from each other.

We also studied the prevalence of functional promiscuity (the ability to perform more than one function) in computational models of the genotype-phenotype map. In particular, we studied RNA, Boolean GRNs and toy-LIFE. Our results suggest that promiscuity is the norm, rather than the exception. These results prompt us to rethink our understanding of biology as a neatly functioning machine. One of the most interesting results of this thesis came from studying the evolutionary dynamics of shifting environments in populations showing functional promiscuity: our results show that there is an optimal frequency of change that minimizes the time to extinction of the population.

Finally, we presented a new metaphor for molecular evolution: adaptive multiscapes. This framework intends to update the fitness landscape metaphor proposed by Sewall Wright in the 1930s. Adaptive multiscapes include many features that we have learned from computational studies of the genotype-phenotype map, and that have been discussed throughout the thesis. The existence of neutral networks, the asymmetry in phenotype sizes -and the concomitant asymmetry in phenotype accessibility- and the presence of functional promiscuity all alter the original fitness landscape picture.

Contents

1

Introduction

"There is more in the diary than just the map."

Henry Jones
Indiana Jones and the Last Crusade (1989)

Evolutionary biology has come a long way since the publication of *On the Origin of Species* by Charles Darwin (1859), more than 150 years ago. The gargantuan work undertaken by Darwin was soon expanded by the contributions of many scientists, helping turn evolutionary biology into the complex discipline that we know today. In the (roughly) half a century after Darwin's book was first published, evolutionary biology received many contributions from foremost scientists: Francis Galton's work on heredity, August Weismann's germplasm theory, Ernst Haeckel's "ontogeny recapitulates phylogeny", the rediscovery of Mendel's laws of inheritance by Hugo de Vries, Carl Correns, Erich von Tschermak and William Jasper Spillman —notably, all four men rediscovered these laws independently—, the origin of genetics with William Bateson or the role of chromosomes in heredity, discovered by Thomas Morgan.

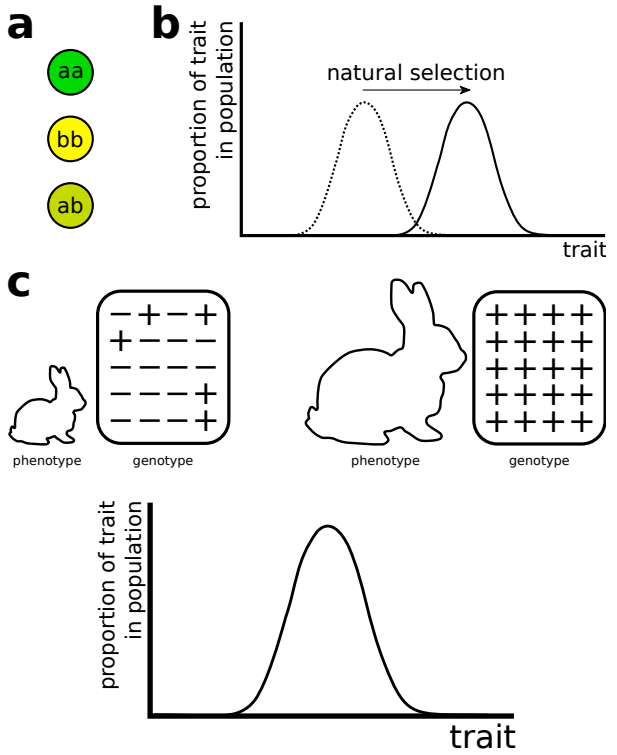


Figure 1.1: Reconciliation between Mendelians and biometricians by Fisher's mathematical models. **a**. If the color of a pea is determined by one gene with two variants (*a* and *b*), then the peas can have at most three different colors, according to Mendel's laws of inheritance. Mendelians thought that inheritance was particulate, and therefore rejected the gradual evolution that natural selection implied. **b**. In the biometricians' studies, traits had continuous, Gaussian distributions in the populations, whose mean and variance would be affected by natural selection. This view was opposite to that of Mendelians. **c**. Fisher proposed a model of the genotype-phenotype map in which many genes would have a small, additive effect on the phenotype. He showed that this combination would yield a Gaussian distribution of the trait in the population, reconciling the Mendelian and biometrician views. In the figure, each gene —represented by a plus or minus sign— contributes a little to the rabbit's size. The final size of the rabbit is just dependent on the total number of plus signs, and not on the identity of the particular genes. This very simple example yields a binomial distribution for the distribution of rabbit's sizes, which is very close to a Gaussian when the number of genes involved in the trait is large.

At the beginning of the XXth century, there were two main schools of thought in evolutionary biology (Bowler, 1989). Mendelism, lead by William Bateson and Hugo de Vries, believed —correctly— in the existence of particulate genes that would be transmitted from one generation to another by Mendel's laws. As these genes were discrete, changes originated in them due to mutations would translate into discrete changes in the phenotype —the collection of observable features in an organism. If the color of a pea is determined by one gene with two variants, then the color that we observe can at most take three different values in a diploid organism —see Figure 1.1**a**. Thus, some Mendelians rejected Darwin's natural selection, arguing that evolutionary changes would proceed through discrete leaps, not continuous variation as Darwin had proposed (Darwin, 1859). On the other side of the argument, biometricians, led by Karl Pearson and Raphael Weldon, studied traits that showed continuous variations, such as height, weight or leg length. They rejected Mendelian inheritance, because discrete units of inheritance could not explain the continuous range of variation found in those aforementioned traits. On the other hand, they embraced Darwin's natural selection, depicting the process of evolution as a change in the distribution of these continuous traits in the population —see Figure 1.1**b** (Bowler, 1989).

The publication in 1930 of *The Genetical Theory of Natural Selection* by Ronald Fisher laid the foundations of population genetics. Using statistical models, he showed that Mendelian inheritance and natural selection, Darwin's proposed mechanism for evolution, were not at odds with each other. Fisher's work served to show that if an organism's features are the product of many discrete genes —inherited following Mendel's laws— combining their effects, then the result would be a continuous range of variation for those features, as measured by the biometricians (Fisher, 1930) (Figure 1.1**c**). Both opposing postures, therefore, were finally reconciled by his work (Bowler, 1989). Mathematically, of course, Fisher's argument is none other than the central limit theorem, that states that the distribution of any random variable that is the sum of a large number of independent random variables —in this case, alleles that each add a small genetic effect to the phenotype— will approach that of a Gaussian variable.

In Fisher's view, selection would act to increase the frequency of advantageous mutations. Assuming that the contribution of each individual

gene to overall reproductive ability, or fitness, was small, Fisher concluded that selection would work slowly and gradually towards ever-increasing fitness (Fisher, 1930; Bowler, 1989). Thanks to Fisher's work, but also to J. B. S. Haldane and Sewall Wright, evolution came to be understood as the change of allele —gene variants— frequencies in the population. Their work set the foundations of the Modern Synthesis, which would dominate evolutionary biology's discourse for the remainder of the XXth century.

The Modern Synthesis, spearheaded by Theodosius Dobzhansky, E. B. Ford, Ernst Mayr, Julian Huxley and G. G. Simpson —among others— in the 1940s, represented the point of maturity of evolutionary biology. It was around this period that Dobzhansky wrote the now famous phrase: "Nothing in biology makes sense, but under the light of evolution". Evolutionary biology would, from then on, influence the thinking of biologists in all other areas of the discipline, from molecular biology to epidemiology, from ecology to physiology. Even more, other disciplines such as computer science or sociology started to look at biologic evolution in search of insights.

As in any scientific endeavor, as time goes on and new information is acquired, old theories are replaced by newer ones. Galton's theory of inheritance was rejected when Mendel's laws were rediscovered. Weismann's germ plasm theory was updated after the discovery of DNA and horizontal gene transfer. Haeckel's recapitulation theory has proven to be wrong.

In the years that have passed since the publication of the books that laid the foundations of the Modern Synthesis —*Genetics and the Origin of Species* (1937) by Dobzhansky, *Systematics and the Origin of Species* (1942), by Mayr, *Evolution: the Modern Synthesis* (1942) by Huxley, and *Tempo and Mode in Evolution* (1944), by Simpson— our understanding of evolution has grown significantly. Most of this is due to our growing understanding of the rest of biological science: microbiology, cellular biology, molecular biology, developmental biology, and so on. As our knowledge of biology grows, so must our models of evolution grow with it. Notable efforts in this respect are the neutral theory of molecular evolution, by Motoo Kimura (1983), the extended synthesis proposed by Massimo Pigliucci and others (Pigliucci and Müller, 2010), or Eugene Koonin's post-modern synthesis (Koonin, 2011).

This thesis is an attempt to contribute to the growth of evolutionary theory.

1.1 Genotype and phenotype: why the map matters

In the early days of genetics, Danish botanist Wilhelm Johannsen coined the terms *genotype* and *phenotype* (Johannsen, 1911). By genotype he meant the set of all genes an organism possessed. As for phenotype, Johannsen says, quite ambiguously,

> All types of organisms, distinguishable by direct inspection or only by finer methods of measuring or description, may be characterized as phenotypes. (Johannsen, 1911)

In time, phenotype would come to refer to the composite of observable features of an organism: morphology, organization, behavior, and so on (Fontana, 2006). With the discovery of DNA, the genotype became synonymous with the information encoded in the chromosomes of an organism. This information is used to build new organisms every generation, and therefore the genotype is responsible for the phenotype.

The distinction between genotype and phenotype provides a comfortable framework for the study of evolution: natural selection acts on those observable characteristics we classify as phenotype, because the reproduction rates depend on them. On the other hand, it is the genotype that is transmitted —inherited— generation after generation, carrying the instructions to generate the phenotype. Mutations act on the genotype, producing new phenotypes, that will be subjected to selection, and so on.

However, as our knowledge of biology has advanced, the distinction between genotype and phenotype has become more and more blurry: the phenotype is the product of a complex process of development carried out by a collection of proteins, organelles and other molecules in a cellular context inherited from the previous generation, that is completed by the information coded in the DNA. The expression of the information coded in the DNA can be altered by this cellular context (Ptashne and Gann, 2002), but also by the modification of histones —proteins surrounding the chromosomes, packing and ordering them— through changes induced by the environment and perceived by the phenotype (Goldberg et al., 2007). As

these modifications are inheritable, they should be considered part of the genotype. In other words, phenotype depends on genotype, which also depends on phenotype. On a different note, the fact that selection can act on the frequency of transposable elements, even when they do not code for any protein (Montgomery et al., 1987), or on the frequency of alleles when segregating in the gametes —meiotic drive (Sandler and Novitski, 1957)— , which *a priori* would seem to be properties of the genotype, suggests that these features should also be part of the phenotype.

Trying to translate that huge complexity into formal models is out of our possibilities at this moment —all the more so when molecular biology is still discovering new cellular and molecular mechanisms which are relevant for evolution. Instead of trying to re-define genotype and phenotype, we will restrict ourselves to very general concepts: organisms develop through a complex biological process, influenced by the environment. Selection then acts on every aspect that can influence reproduction, and some information is transmitted to the organism's offspring. The exploration of very simple models of the genotype-phenotype map, as we will call the relationship between these two concepts, will yield insights into the evolutionary process that were not present in the Modern Synthesis.

Because they developed their models in the first half of the XXth century, Fisher, Haldane, Wright and many of the population geneticists that came after them had no knowledge of all these facts, and assumed a simple relationship between the genotype and the phenotype (Fontana, 2006). As we have seen (Figure 1.1), Fisher assumed that a large number of genes would contribute to the phenotype in an additive way, thus giving rise to the Gaussian distribution of characters that the biometricians observed in their measurements. The population geneticists' approach was very useful in that it allowed for a powerful framework that could predict the evolution of allele frequencies given their contributions to fitness. However, the models of population genetics fail to explain many evolutionary phenomena. Among these are punctuated equilibria (Eldredge and Gould, 1972; Gould and Eldredge, 1977), the constraints to evolution (Maynard Smith et al., 1985), or the origins of novelty (Wagner, 2011; Nei, 2013). All these limitations have something in common: they are related to a lack of knowledge regarding the stability and accessibility of phenotypes. More than 20 years ago, Pere Alberch (1991) explained that studying the complexities

of the genotype-phenotype map would lead to the study of new properties that had not been considered relevant before: robustness —how stable phenotypes are regarding mutations and environmental perturbations— and evolvability —how easy it is to reach new phenotypes. In other words, studying the genotype-phenotype map not only solves previously puzzling problems, but also yields new insights into evolution.

Perhaps one particular example will clarify how the introduction of more complexities of the genotype-phenotype map may help to correctly predict evolutionary phenomena. As part of his neutral theory of molecular evolution, Kimura had proposed that the rate of substitution of amino acids in proteins was a random, Poissonian process (Kimura, 1983). That implies that the ratio between the variance and the mean, called the dispersion index, was 1. However, experimental data showed that the dispersion index was greater than 1 (Manrubia and Cuesta, 2015), which would imply that Kimura's neutral hypothesis was wrong (Bastolla et al., 1999). Using population genetics, Gillespie (1991) argued that natural selection had to be invoked in order to explain the data. That is not necessarily the case. Recent studies of the genotype-phenotype map have shown that genotypes form heterogeneous neutral networks (see next section): different genotypes have a different number of neutral neighbors, affecting the substitution rate. When taking into account this fact, the resulting stochastic process of substitution is not Poissonian anymore, accounting for the high values of the dispersion index (Bastolla et al., 1999; Manrubia and Cuesta, 2015), and eliminating the need for natural selection.

1.2 Surveying the genotype-phenotype map

Throughout this section we will restrict our study to unicellular organisms. The phenotype of multicellular organisms is, of course, much more complex and it involves many features like developmental processes, tissue differentiation, apoptosis and nervous and hormonal control systems. All of these features no doubt increase the potential for evolvability, as they increase the dimension of the search space, but they escape the scope of this introduction —see, however, Chapter 5 for a brief detour into multicellular phenotypes.

Very briefly, we can outline the components of the genotype-phenotype map in four big blocks: (i) the storage of hereditary information in nucleic acids; (ii) the transcription and translation of this information into "effector molecules" —RNA and proteins, which will carry out most functions in the cell; (iii) regulatory networks, that decide when a piece of information will be expressed and (iv) metabolic networks, which make it possible for a cell to gather energy in order to grow and divide, the ultimate goal of any living being. It is a powerful argument for the common ancestry of all living beings that most organisms share the same molecule to store information, use the same code to translate that information into the same effector molecules, and share regulatory and metabolic components.

The first block is what is usually termed the genotype, as we have already discussed. What we call the phenotype is a combination of the other three blocks. So far we lack the means to study the relationship between the four blocks in a comprehensive way.

Take, for example, *Escherichia coli*. *E. coli* is a unicellular bacterium, and is one of the best well-known organisms in the planet. Its genome has been sequenced, with around 4 million base pairs long, comprising around 4,000 genes (Rudd, 2000). Most of that genome has been annotated, which means we know which products are expressed by different parts. There is much information regarding its regulatory network (Gama-Castro et al., 2015). *E. coli* has been cultured in many different media, giving us insight into its metabolic capabilities (Orth et al., 2011). It has been the subject of several experimental evolution studies, some of them long-term (Lenski et al., 1991; Elena and Lenski, 2003). And yet, when a single new mutation occurs, it is almost impossible to know how it will affect the phenotype of the bacterium.

Consider, for instance, the *tauA* gene. This gene codes for a protein that forms part of a taurine transporter, a multimeric compound that takes this amino acid into the cell. We can predict the structure of the protein with some accuracy, thanks to advances in protein folding algorithms (see Figure 1.2 and Section 1.2.2). Suppose now that the gene suffers a substitution mutation in an arbitrary position. Our prediction algorithms for protein folding allow us to predict changes in the tertiary structure. However, we do not know what effect this particular mutation will have on the phenotype. How does this change affect the assembly of the transporter?

Figure 1.2: Tertiary structure of tauA. tauA is part of a complex that transports the amino acid taurine into *E. coli* cells. Although we can predict its structure with some accuracy, we cannot know which will be the effect of a mutation on its function. Structure obtained from `proteinmodelportal.org` (Haas et al., 2013).

Will it increase taurine transport? Will it decrease it? Will nothing change noticeably? Without any experiments performed to answer these particular questions, we can say nothing.

And this is just one gene among the 4,000 that comprise the genome of one of our most well-known model organisms, only considering the effect of point mutations. Consider how complex the picture becomes if we try to introduce the effect of duplications, inversions, and so on.

As a result, because this endeavor is still out of our possibilities, some researchers have turned their attention to computational models of the genotype-phenotype map in which both the genotype and the phenotype are greatly simplified — a simplification that enables exhaustive computational explorations while at the same time extracting useful knowledge from them.

As for the genotype, models of the genotype-phenotype map usually assume it to be just a string of letters belonging to an alphabet —either the four bases of DNA or a binary one. Mutations, or changes in the genotype, are usually just point mutations, in which one of these letters in the string is changed for another one. The complexities of the storage of information are thus ignored: no attention is paid to the storage of information in

histones, the proteins surrounding and compacting the DNA into chromosomes, in what has come to be termed epigenetics (Goldberg et al., 2007). Moreover, the complex processes behind the appearance of mutations are also ignored, such as different rates for different parts of the genome —the so-called hotspots of mutation—, the frequency of insertions and deletions, and so on.

Regarding the phenotype, these models usually focus on just one component —RNA, proteins— or on the structure of a network —regulatory and metabolic networks. In general, every model presents a rule that maps strings of letters into elements of phenotype space, which we will describe in the next sections.

Although simplistic and far from the complexities of the real genotype-phenotype map, these studies try to understand what kind of phenotypic effect different mutations will have, and what is the general structure of the genotype-phenotype maps, trying to see if there are any universal properties underlying all of them. Along this thesis we will see that there are indeed some lessons to be extracted from these simplified toy models.

1.2.1 RNA

DNA carries (almost) all the hereditary information, but it cannot do anything with this information, except copy it. In order for this information to build a cell and carry out all cellular functions, DNA must be expressed. That is, it must be transferred to molecules that can actually do something with it. These molecules are RNA and proteins, the key effector molecules in all cells.

RNA molecules are generated from sequences of DNA that are used as templates in a process called transcription (Alberts et al., 2014). Molecular biologists classify RNA molecules according to their functions, and consequently we have:

mRNAs: messenger RNAs. They are translated into proteins (see below). mRNAs are very similar to DNA in that they only store information. They are very different, though, in that they are much more fragile than DNA: the half-life of a mRNA molecule can vary from minutes in some bacteria, to hours in mammalian cells (Milo et al., 2010). When transcription occurs, multiple mRNAs are copied at the same

time: as much as 1,000 transcripts can be synthesized in one hour from a single gene.

rRNAs: ribosomal RNAs. They form the basic structure of the ribosome, which is the essential macro-polymer that catalyzes protein synthesis. The number of ribosomes goes from $10^4 - 10^5$ in bacteria to $10^6 - 10^7$ in eukaryotic cells (Milo et al., 2010).

tRNAs: transfer RNAs. They function as adaptors between mRNA and amino acids during protein synthesis.

miRNAs: microRNAs. They regulate gene expression.

There are many more types of RNA, most of them with regulatory or scaffolding functions. Because of their single-stranded nature, RNA molecules are more flexible than DNA and can fold into a three-dimensional structure, called the tertiary structure. This shape enables the RNA to perform its function. The tertiary structure is heavily conditioned by a two-dimensional structure called the secondary structure (Huynen, 1996; Aguirre et al., 2011). Because the tertiary structure is correlated with function, the secondary structure of an RNA sequence is often considered a good proxy for its phenotype (see Figure 1.3).

RNA sequences, their secondary structures, and the folding from the former to the latter are, by far, the most studied of the computational genotype-phenotype maps. In the mid 1990s the Vienna group, led by Peter Schuster, developed a computational algorithm to predict the secondary structure of RNA molecules (Hofacker et al., 1994), based on thermodynamic principles and biochemical restrictions.

The fast prediction algorithm developed allowed the Vienna group to study exhaustively the properties of the mapping from RNA sequences to secondary structures (Schuster et al., 1994; Grüner et al., 1996a; Fontana and Schuster, 1998), and it has been used by many other research groups since then (Jörg et al., 2008; Aguirre et al., 2011; Dingle et al., 2015). As a result, this particular genotype-phenotype map is very well characterized.

The Vienna group was the first to fully map a complete set of sequences to their corresponding structures, and study the associated statistics (Schuster et al., 1994), obtaining the results we now associate with most computational genotype-phenotype maps. Thus, for example, they found that, for

a fixed sequence length L, the number of theoretically possible structures is much smaller than the total number of sequences. Moreover, after folding all sequences into their corresponding structures, many theoretically possible structures were actually not mapped by any sequence. That is because many sequences fold into the same structure. This means, in other words, that the genotype-phenotype map is highly degenerate. In RNA, it has been calculated (Schuster et al., 1994) that the number of secondary structures scales as $L^{-3/2}1.8^L$, while the number of sequences grows as 4^L. This implies that the average number of sequences that fold into a given structure grows as $L^{3/2}2.16^L$, which is an enormous number, even for moderate values of L.

However, not all structures are equally frequent —i.e. not all of them are mapped by the same number of sequences. In fact, the distribution of frequencies is highly skewed. If we order the structures according to

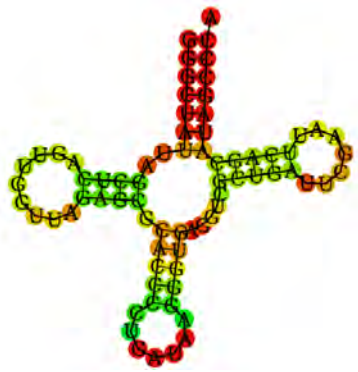


Figure 1.3: Secondary structure of a sample RNA molecule. Output of the RNA folding algorithm developed by the Vienna group as implemented in the RNAfold server (Hofacker, 2003). The folded sequence is GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCCUGAUAAGGGUGAGGUCGCUGAUUCGAAUUCAGCAUAGCCCA, a sample sequence provided by the server. The secondary structure can also be represented in a compact way using the dot-bracket notation (Aguirre et al., 2011): an opening parenthesis (represents a base which is paired with a nucleotide closer to the 3' end, a closing parenthesis) a base paired with a nucleotide closer to the 5' end, and dots denote unpaired nucleotides. Using that notation, the structure is (((((((..((((.....)))).(((((.....)))))).....(((((.....)))))))))))). .

their frequency, and plot these frequencies with respect to the rank of the structure, we obtain a power-law distribution (Schuster et al., 1994). This means that most structures are rare —in the sense that only a few sequences fold into them— and only a few structures are very common. Moreover, the probability density function associated with the frequencies of those structures is a log-normal distribution (Dingle et al., 2015).

As sequence length increases, this skewness seems to increase: that is, most sequences fold into a decreasing proportion of structures. This implies that small phenotypes will not play a central role in evolution: they are hard to find in a genotype space that is filled with abundant structures (Schaper and Louis, 2014). In fact, it has been described that RNA molecules found in nature correspond only to those highly frequent structures in sequence space (Jörg et al., 2008; Dingle et al., 2015).

For a fixed sequence length L, we can assign a topology to the space of sequences. Two sequences will be connected if they differ in only one nucleotide —i.e. a point mutation. The whole of sequence —or genotype— space will then become a regular undirected graph, in which every genotype will have exactly $3L$ neighbors, in the case of RNA. The folding of those sequences into their corresponding secondary structures colors this regular graph, partitioning the genotype space into the set of attainable structures, or phenotypes. Because the number of phenotypes is much smaller than the total number of genotypes, a given sequence will typically be surrounded by neighbors that fold into the same phenotype. The set of genotypes that fold into a given phenotype will be organized into a network, called *neutral network*. Normally, these neutral networks form a giant connected component, although this is not always the case (Aguirre et al., 2011).

The degree of a node in this neutral network, that is, the number of neighbors it has, is usually called the *genotypic robustness* of the node. It is usually normalized by the total amount of neighbors in genotype space, representing then the fraction of neighbors that share the same phenotype (Wagner, 2011). Inside a particular neutral network, the degree distribution is highly heterogeneous. Aguirre et al. (2011) find that, for sequences of length $L = 12$, the distribution is wide, with one peak. Wagner (2011) also states that degree distributions in RNA are wide, with a mode on $0.2 - 0.3$ (relative to maximum degree). The average degree of a neutral network is

related to its size (Aguirre et al., 2011): larger networks have greater average degrees, and this relationship is logarithmic —that is, average degree grows with the logarithm of network size.

The topology on the sequence space can be complemented with a metric, normally using Hamming distance, which measures the number of positions in which two sequences are different. Using that metric, Wagner (2011) performed neutral walks along a neutral network —that is, each time step the genotype is mutated, and the mutation is accepted only if the phenotype is not changed. He also forced the walk to increase the Hamming distance every step, and found that the average final distance to the original genotype was very close to the maximum. This means that RNA neutral networks percolate the whole sequence space, and that they contain sequences that fold into the same structure but that don't necessarily share any base (Schuster et al., 1994; Wagner, 2011). Conversely, if we take one sequence and change around 15% of it at random, the resulting sequence will fold into a structure that will be as similar to the original one as if we took a structure at random from the set of possible ones (Huynen et al., 1993).

That is, not only most sequences have neighbors that share the same phenotype as them, but also around any sequence there is a small neighborhood that contains all common phenotypes (Grüner et al., 1996b; Wagner, 2011). The minimum radius that needs to be explored around a given sequence in order to find the most common structures is very small, which means that most common phenotypes are close to each other. For instance, for RNAs of length L=100, a sphere of radius 15 contains with probability one a sequence for any frequent structure (Wagner, 2011). The number of sequences contained in this radius is infinitesimally small compared with the total number of sequences in genotype space. This means that the neutral networks belonging to common phenotypes are intertwined between each other. This phenomenon is what Schuster and collaborators termed *shape space covering* (Schuster et al., 1994; Grüner et al., 1996b). This means that those phenotypes that are more abundant are easily accessible from any other phenotype, so that the search for new structures among frequent ones is a fast evolutionary process.

Finally, inside a neutral network, the phenotypes associated to a neighborhood change. Performing a neutral walk along a RNA neutral network

—without forcing it to increase distance every step— the cumulative number of phenotypes belonging to neighborhoods of genotypes visited during the random walk increases linearly (Huynen, 1996). Also, the similarity between the neighborhoods of genotypes across the neutral network decreases with the distance among sequences (Huynen, 1996; Wagner, 2011).

The features of RNA neutral networks are not unique of this particular genotype-phenotype map, as we will see.

1.2.2 Proteins

Although there are RNA molecules with catalytic activity (ribozymes, like rRNA), most of the actual functions in a cell are carried out by proteins: transport, metabolism, DNA replication, RNA transcription, and so on. Proteins are long, unbranched polymer chains, formed by a string of amino acids, chosen from an alphabet of 20 types of amino acids. Proteins are built from the information encoded in mRNAs, in a process called translation. The code the cell uses to transform the nucleotide alphabet into an amino acid alphabet is called the genetic code and, apart from some minor exceptions, it is universal among all living beings. The genetic code maps every possible sequence of three nucleotides into an amino acid. There are 4 nucleotides, so there are $4^3 = 64$ possible sequences of three nucleotides. As there are only 20 amino acids, this means that some sequences must map into the same amino acid. We say that the genetic code is degenerate. The sequence of nucleotides is read by tRNAs: each one of them becomes attached to one amino acid at one end, and to the three-base sequence at the other end. The three base sequence in the mRNA is called the codon, and its complement in the tRNA is called the anticodon. There is at least one —typically many more— tRNA for each anticodon. The mRNA molecule goes through the ribosome, which binds together the different amino acids that the tRNAs carry, to form a full protein. The points where a mRNA must start and end its translation are marked on the same mRNA molecule: there are “start” and “stop” codons (Alberts et al., 2014).

Once formed, the amino acid sequence starts to fold into a (generally) compact three-dimensional structure called the tertiary structure, that will allow the protein to perform its function. Unlike the folding of RNA into a secondary structure, which is well studied and characterized (see previous

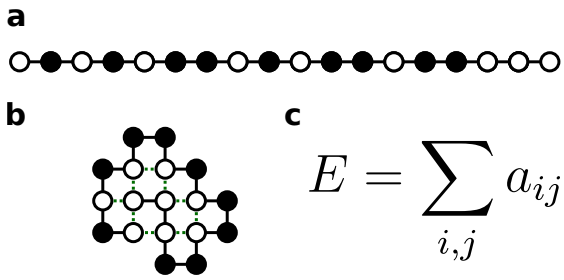


Figure 1.4: Protein folding in the HP model. a. In the HP model, the genotype of a protein is just a sequence of H (white circles) and P (black circles) amino acids. **b**. The protein folds into a discrete lattice, minimizing its free energy. **c**. The free energy of a folded protein is computed as the sum of the interacting energies a_{ij} between adjacent amino acids —marked in green in **b**.

section), the protein folding problem has not been solved yet. In fact, it is known to be NP-complete, which means that there are no fast algorithms to predict the tertiary structure associated to a sequence (Berger and Leighton, 1998).

The function of proteins is related to its fold, but not determined by it. Enzymes perform their function through small sites in their structure, so many different enzymes can have very similar folds and still perform very different functions (Wagner, 2011). A paradigm for this is hemoglobin S, the protein responsible for sickle-cell disease. Hemoglobin S is just a point mutation away from hemoglobin A, the hemoglobin that appears in healthy red blood cells. The only difference between the two proteins is an amino acid. The tertiary structure, studied through crystallography, is almost identical in both proteins. However, this single amino acid change alters the way different hemoglobin molecules interact with each other, causing a polymerization of hemoglobin S into long fibers that disrupt the form of the red blood cell and cause disease (Eaton and Hofrichter, 1990). Identical tertiary structures lead to very different functional outcomes.

These considerations notwithstanding, most studies of the genotype-phenotype map focusing on proteins are centered on the tertiary structure, in particular through lattice models (Wagner, 2011). Lattice models force the sequence of amino acids to fold into a discrete lattice, that can be two or three-dimensional. Perhaps the most well-known of these models is the

HP model proposed by Lau and Dill (1989). In the HP model, proteins are formed by strings of two amino acids: hydrophobic (H) and polar (P). The sequence will try to fold into a compact structure, minimizing its free energy. The original model assumed that the only way to decrease free energy was through interaction between H residues (Lau and Dill, 1989) —two residues interact if they are adjacent in the lattice, but not in the sequence (see Figure 1.4). Subsequent modifications have included interaction energies between H and P residues as well (Li et al., 1996). The idea behind this algorithm is that, in water, proteins will try to hide their hydrophobic residues inside their core, forming a compact structure. Thus, the genotype considered in this model is just a binary string of varying length, while the phenotype is the structure obtained in the lattice. Note that focusing on the sequence of amino acids as the genotype eliminates all the complexities of transcription and translation from the map.

When a protein sequence folds into a unique structure, it is usually termed *designing sequence* (Li et al., 1996; Irbäck and Troein, 2002). Proteins that do not fold into a unique structure are usually not considered (Li et al., 1996; Irbäck and Troein, 2002).

Similarly to what we saw in RNA, the HP model generates a small number of structures compared with the total number of sequences — although there is an important difference in that the number of protein sequences that do not fold into a unique structure is much higher (Ferrada and Wagner, 2012). Irbäck and Troein (2002) folded all designing sequences up to length 25 and found that, for example, there are $765,147$ designing sequences of length 25 —only 2.28% of all possible sequences— that fold into $107,336$ structures. Li et al. (1996) folded all sequences of length 27 into a compact 3×3 cube, and found that the number of structures was $51,704$ —although many of them were not the result of the folding of any sequence, while the number of designing sequences was 4.75% of all $2^{27} \sim 10^8$ possible sequences. Bornberg-Bauer (1997) folded all sequences of length 18, and found that, out of the $2^{18} = 262,144$ possible sequences, only 2.4% are designing sequences, and fold into $1,475$ structures. As we can see, the ratio of sequences to structures is not as high as in the case of RNA: protein sequences in the HP model generate more different structures, and therefore the average number of sequences that fold into a given structure is lower than in the case of RNA (Ferrada and Wagner, 2012).

Again, not all these structures are equally frequent. Li et al (Li et al., 1996) find that some structures are highly frequent, while most are rare. Also they describe those frequent structures to be "protein-like", implying that structures that appear frequently in cells are just the most frequent in phenotype space —again showing that the constraints imposed by the genotype-phenotype map have a strong effect on evolution. The frequency distribution of structures is close to an exponential distribution (Li et al., 1996). Bornberg-bauer (Bornberg-Bauer, 1997) found that, plotting the rank-ordered structures according to their frequency —as we saw in RNA— the distribution was also a power law.

The HP model also produces neutral networks of protein structures (Lipman and Wilbur, 1991; Li et al., 1996; Bornberg-Bauer, 1997; Bastolla et al., 1999). This is in accord with experimental and computational data showing that proteins sharing the same fold need not share the same sequence (Babajide et al., 1997; Rost et al., 1998). Moreover, most of these neutral networks form one connected component (Bornberg-Bauer, 1997).

So far, the similarities of the HP model and the RNA genotype-phenotype map are striking. There are, however, two main differences between the results of the two models (Bornberg-Bauer, 1997; Ferrada and Wagner, 2012). In RNA, neutral networks percolated through sequence space, while in the HP model, neutral networks are clustered in sequence space, around some prototypical structures that have a high amount of neutral neighbors. Secondly, in the HP model there is no shape space covering —that is, it is difficult to transform one common structure into another (Ferrada and Wagner, 2012; Bornberg-Bauer, 1997). This does not imply, however, that there is no diversity in the neighborhood of a given neutral network: as was the case with RNA, different protein sequences inside a neutral network have different structures as neighbors (Wagner, 2011; Ferrada and Wagner, 2012).

1.2.3 Regulatory networks

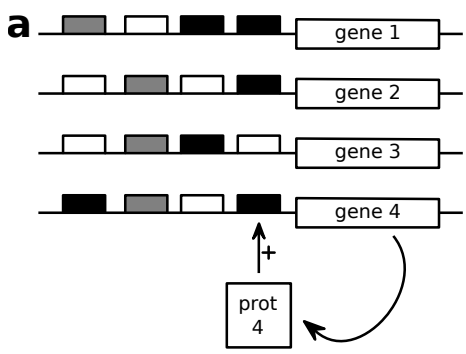
Genes are transcribed into mRNAs which, in turn, are translated into proteins. The transcription of genes is started by the RNA polymerase, and some proteins have the ability to either enhance or inhibit the activity of the polymerase. These proteins are called transcription factors (Alberts et al.,

2014). It is easy to see how cells can develop complex expression patterns by producing proteins that will either enhance or inhibit the expression of other proteins. This regulation of expression is even more complex when we take into account that the effect of transcription factors can be modulated by other proteins that do not interact directly with the polymerase, and the modulation imposed by signals perceived from the exterior of the cell (Ptashne and Gann, 2002). All of these proteins form a complex regulatory network, that will enable the cells to respond to environmental stimuli, expressing those genes that are fundamental to the problem at hand. Regulatory networks are also responsible for the cyclic expression of genes —such as those affecting the cell cycle of division (Alberts et al., 2014).

Modelling of gene regulatory networks (GRNs) usually focuses on the spatio-temporal patterns of gene expression using differential equations (see Garcia-Ojalvo (2011) and Rué and Garcia-Ojalvo (2013) for reviews). However, the number of circuits that can be designed with just three genes is astronomical, and if we consider continuous variations of interaction parameters, the number is infinite. Therefore, the study of the genotype-phenotype map restricted to regulatory networks has been mostly devoted to discrete, Boolean networks — but see Schaerli et al. (2014) and Jiménez et al. (2015) for some recent work on continuous regulatory networks.

In discrete Boolean networks, time is discrete, and at each time step, genes can be either ON or OFF. A given expression profile at a certain time step determines —if the network is deterministic— the genes that will be expressed at the next time step: the gene products are assumed to affect the expression of other genes. Thus, the state of a Boolean network at time $t+1$ only depends on the state of the network at time t. Originally proposed by Stuart Kauffman (1969), Boolean networks have been used to describe the GRN of yeast (Kauffman et al., 2003), fruit fly (Albert and Othmer, 2003) and *Arabidopsis thaliana* (Espinosa-Soto et al., 2004).

The phenotype of a Boolean regulatory network is taken to be the temporal expression pattern of the genes that belong to it. All the complex processes behind the functioning of regulatory networks —transcription of RNAs, translation of proteins and interaction between proteins— are difficult to include in a computational model that allows for exhaustive exploration of genotype space. As a result, models of the genotype-phenotype map that focus on regulatory networks take the genotype to be the net-



b

input	output
0 0 0	1 0 0
0 0 1	0 1 1
0 1 0	0 0 1
0 1 1	1 1 1
1 0 0	1 0 1
1 0 1	0 1 1
1 1 0	0 1 1
1 1 1	1 0 1

Figure 1.5: Boolean regulatory networks. a. Wagner's original model (Wagner, 1996) describes regulatory regions as a combination of sites where transcription factors will bind, enhancing or inhibiting the interaction of the corresponding gene. In this figure, the regulatory region is formed by four genes, each with four sites where transcriptional factors can bind. The effect of each transcription factor can be activating (black), null (white) or inhibiting (gray) regarding gene expression. As a guide, a picture of the gene product of gene 4 is drawn —called prot 4—, and we show the positive interaction it has on its own expression. The final expression state of a given gene at time t will be the result of the combination of the inhibitory and activating effects of the gene products of all four genes in the network, in an additive way. See text for more details. **b**. In Payne et al's model (Payne et al., 2014), the genotype is the output expression of every gene at time $t+1$, given all possible input states. This figure represents a network formed by three genes. When, at time t, all genes are active —last row on the input column—, the state at time $t+1$ will express the products of genes 1 and 3, but not of gene 2. This genotype can be thought of as the combination of the three truth tables associated with each gene, and has the advantage that it can represent non-additive Boolean functions.

work of interactions between genes: which genes are activated or inhibited by each of the genes belonging to the network. Most of this work has been done by Andreas Wagner and collaborators (Wagner, 2011), and we will briefly review some of the main results of two of their computational models, that differ in the way they map genotypes to phenotypes. Both of them, however, assume that transcription factors (proteins) evolve much more slowly than the binding sites that affect transcription —also called *cis*-regulatory elements.

In the first model, developed by Wagner (1996), the genotype is a matrix $\mathbf{W}$ whose element w_{ij} describes the effect that gene i has on gene j: activation ($w_{ij} > 0$), no effect ($w_{ij} = 0$) or inhibition ($w_{ij} < 0$). The state of gene i at time $t+1$, $E_i(t+1)$, will be given by

$$E_i(t+1) = \Theta\left(\sum_j w_{ij} E_j(t)\right), \tag{1.1}$$

where $\Theta(x)$ is the Heaviside function ($\Theta(x) = 1$ if $x > 0$ and $\Theta(x) = 0$ otherwise). In most works, the values of w_{ij} are chosen from the set $\{+1, 0, -1\}$, thus simplifying the regulatory effect to activation, no effect or inhibition, respectively (Wagner, 2011) (see Figure 1.5**a** for a schematic of what this genotype definition means in biological terms). The temporal pattern of gene expression is computed starting from an arbitrary initial state. Usually only interaction matrices that produce fixed temporal patterns are considered as viable phenotypes (Ciliberti et al., 2007a; Wagner, 2011).

Two genotypes are connected in genotype space if their interaction matrices W and W' differ only in one element. In other words, the model considers as potential mutations those that only alter the interaction between two genes (Ciliberti et al., 2007a; Wagner, 2011).

In this model, there are exponentially many more genotypes than phenotypes, which means that many genotypes share the same expression patterns. Again, some of these expression patterns are common, while most of them are rare (Wagner, 2011).

Regulatory genotypes form neutral networks that percolate sequence space (Ciliberti et al., 2007a): when sampling pairs of interaction matrices that express the same phenotype, Ciliberti et al (Ciliberti et al., 2007a) found that the average Hamming distance between them is close to 0.8 relative to the maximum distance. Performing neutral random walks in genotype space, forcing them to increase the distance to the original genotype, they found that the distance could grow as high as 1. These networks form usually one very large component (Wagner, 2011), with perhaps some small fraction of genotypes belonging to independent components.

As for the distribution of neutral neighbors for genotypes in a neutral network, it is unimodal and very broad, even more so than in RNA (Ciliberti et al., 2007b; Wagner, 2011).

The diversity of neighborhoods associated to genotypes belonging to the same neutral network is also very high, as was the case in RNA and the HP model: the phenotypes that appear in the neighborhoods of two genotypes belonging to the same neutral network become more different as the genotypic distance between the two genotypes increases (Wagner, 2011).

As was the case in RNA, neutral networks belonging to two different phenotypes are close to each other in genotype space (Wagner, 2011): the minimal genotypic distance separating two phenotypes is close to 0.15 of the maximum in regulatory networks formed by 20 genes and 5 interactions per gene on average, and never higher than 0.3 (Wagner, 2011).

The second model studied by Joshua Payne in Wagner's group (Payne et al., 2014) defines the genotype as the truth table that assigns an output state (the state in time $t+1$) given each possible combination of expression profiles in time t (see Figure 1.5**b**). In a network composed by n genes, there are 2^n expression profiles that can act as input states. Each of these input states will result in an output state for each of the n genes. There are 2^n output states for each of the 2^n input states, so the total number of genotypes is $(2^n)^{(2^n)}$. However, the number of phenotypes —expression patterns— is much smaller. Combinatorial analyses show that the number of possible expression patterns grows more slowly than the number of possible genotypes (see Appendix A.2), and many of these are not even the result of the expression of any genotype. Payne et al. (2014) show how this model again generates neutral networks that traverse genotype space, with diverse neighborhoods and broad distributions of neutral neighbors.

1.2.4 Metabolism

At each moment in time, the cell must face different challenges from the environment that surrounds it. Food in the form of different chemicals, toxic compounds, signals from neighboring cells, and so on. In order to respond to these challenges, the cell must combine the expression of different proteins at different times, combining all their different functions. We will focus here on metabolic networks, that is, the networks formed by all the enzymes a cell possess.

Enzymes are proteins that catalyze chemical reactions, transforming agents into other compounds that the cell can use to build new components, or to get energy, for example. Most reaction pathways make use of more than one enzyme. And, conversely, many enzymes form part of different pathways (Wagner, 2011). Additionally, *E. coli* and yeast are known to be able to perform more metabolic reactions than are actually needed in a given environment (Rodrigues and Wagner, 2009; Wang and Zhang, 2009; Güell et al., 2014). Metabolic networks, as a result, are complex entities, whose evolution is highly important for the survival of cell lineages.

Again, studying the genotype-phenotype map regarding metabolic networks is impossible if we try to take into account the complexities of transcription and translation rates, protein interactions, and so on. In order to perform exhaustive computational analyses, Wagner's group (Wagner, 2011) has focused on genotypes formed by metabolic reactions. Representing all metabolic reactions as a list, a metabolic genotype is a vector whose i-th component is 1 if the i-the reaction is present in the cell genome, and 0 otherwise. Using flux balance analysis, they study under which carbon (Rodrigues and Wagner, 2009; Wagner, 2011; Hosseini et al., 2015), sulfur (Rodrigues and Wagner, 2011) or nitrogen (Wagner et al., 2014) source the given genotype is able to synthesize all essential biomass molecules in a given environment. In other words, for each sub-metabolism (carbon, sulfur, nitrogen), the phenotype will be a binary vector containing a 1 in the i-th position if the cell is able to survive in the presence of the i-th metabolic source alone. In all cases, they find the same characteristics described before for previous models of the genotype-phenotype map: genotypes form neutral networks of the same phenotype, these networks traverse genotype space, the distance between different phenotypes is small, and genotypes belonging to the same neutral network have different neighborhoods.

1.3 Summary

Maynard Smith (1970) already proposed that, for evolution to occur, neutral networks have to exist: naturally occurring proteins should have mutational neighbors that have some viability as well. Years after his prediction, neutral networks have been confirmed to exist by both experimental and

computational analyses. Additionally, these neutral networks have many properties that enable and facilitate adaptation and innovation.

It is somewhat striking that computational models as different as the folding of RNA, the HP model, regulatory networks and metabolism show so many similarities. Other computational models, such as the polyomino model (Johnston et al., 2011; Greenbury et al., 2014) that we have not explored here, also share most of these properties. Let us summarize them here:

- Degeneracy: many genotypes generate the same phenotype.
- Skewness in frequencies: the distribution of abundances of phenotypes is highly skewed: there are a few very frequent phenotypes, while most of them are rare.
- Neutrality: the genotypes expressing the same phenotype usually form large neutral networks in which genotypes have a broad distribution of neutral neighbors. Typically, these neutral networks traverse genotype space, meaning that they contain very diverse genotypes.
- Accessibility: most common phenotypes are accessible through every other common phenotype by mutations. It is also related to what the Vienna group termed shape space covering: around a given genotype there is a small radius of mutations inside which every common phenotype is found. The HP model does not share this property, however, possibly because considering only compact proteins is too restrictive.

We should note that, although neutral networks can be large and we refer to some phenotypes as “frequent”, in fact even the most common phenotypes represent very small fractions of genotype space (Schuster et al., 1994; Wagner, 2011).

1.4 Objectives

Fisher’s assumption of a simple relationship between the genotype and the phenotype allowed him to explain how natural selection could work in a

gradual fashion with particulate inheritance. The acknowledgment of neutral networks explains the over-dispersion of the substitution rate in proteins. In order for our understanding of evolution to advance, we need to keep delving into the complexities of the genotype-phenotype map.

In this thesis, we will propose a new model of the genotype-phenotype map that includes several levels of expression in a single model. This model, called toyLIFE, includes a simplification of genes, proteins and metabolites, as well as the processes of translation, regulation and metabolism. The main idea behind toyLIFE is to study if including the different levels in the same model has any effect over the main properties described in Section 1.3.

Chapter 2 will be devoted to the definition of toyLIFE, and Chapter 3 to the exploration of the metabolic genotype-phenotype map that it generates.

The capabilities of toyLIFE are not restricted to studying the properties of the genotype-phenotype map, however. It also generates intuition into interesting evolutionary phenomena, such as functional promiscuity —the ability of molecular phenotypes to perform more than one function. In Chapter 4 we will explore how toyLIFE is a good model to study functional promiscuity, and study some of its dynamical consequences for evolution.

In Chapter 5 we will briefly explore a simple multicellular phenotype in toyLIFE, using only its regulatory aspects. We will explore the spatio-temporal patterns of gene expression when a one-dimensional array of cells is studied.

Chapter 6 will summarize some of the lessons learned from all models of the genotype-phenotype map —including toyLIFE— into a new framework that, we hope, helps to better visualize and understand evolution, and that we call *adaptive multiscapes*.

Finally, Chapter 7 will summarize the results obtained in this thesis, and propose some possible lines of research that will advance the topics here studied.

2

toyLIFE

"The chemistry must be respected."

Walter White
Breaking Bad, season 3, episode 5 (2010)

The genotype-phenotype map in real cells is highly complex: genes are transcribed into RNA, which is in turn translated into proteins, which interact with each other and with genes to regulate their expression, thus forming complex regulatory networks. Protein complexes also interact with the environment, performing metabolic reactions that fuel the cell and that create its building blocks. Most models of the genotype-phenotype map, however, focus only on one aspect of this map. Low-level models, such as RNA or protein secondary structure, focus on the biophysical constraints that govern sequence evolution, but leave out the cellular context in which these structures will act, thus losing some complexities associated to function. High-level models, such as gene regulatory networks (GRNs), focus on the interactions between individual components in the cell, thus giving

some insight into this complex cellular context. However, the effect of mutations on function is lost: we do not know how a change in DNA sequence will affect the expression pattern of genes in the cell (see Figure 2.1).

Of course, trying to integrate all this complexity is well out of our current possibilities and knowledge. This is why we have designed toyLIFE, a model that connects the low-level biophysical constraints on sequence-structure relations to the high-level interactions between cellular components. toyLIFE is a very simplified model of a complex genotype-phenotype map and, as such, it does not pretend to mimic cellular biology. The main idea behind the model is to understand if all the properties described for other models of the genotype-phenotype map are maintained when the mapping is more complex. Let us now define toyLIFE.

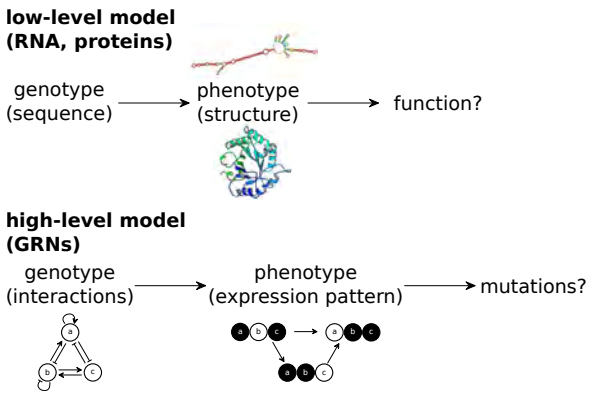


Figure 2.1: Complexity of the genotype-phenotype map. Most models of the genotype-phenotype map take into account one aspect of this map. Low-level models of the map, such as RNA and protein folding, focus on the biophysical constraints on sequence evolution, but leave out the cellular context in which these molecules will perform their function, thus losing meaningful ways in which to assign function. High-level models of the map, such as gene regulatory networks (GRNs), focus on the cellular context, thus incorporating the complexities of function into the definition of phenotype, but leaving out the effects of mutation at the DNA level on these phenotypes. In this case, the genotype are the interactions between genes, and the phenotype is the temporal expression pattern of the genes in the network (black means “OFF” and white means “ON”).

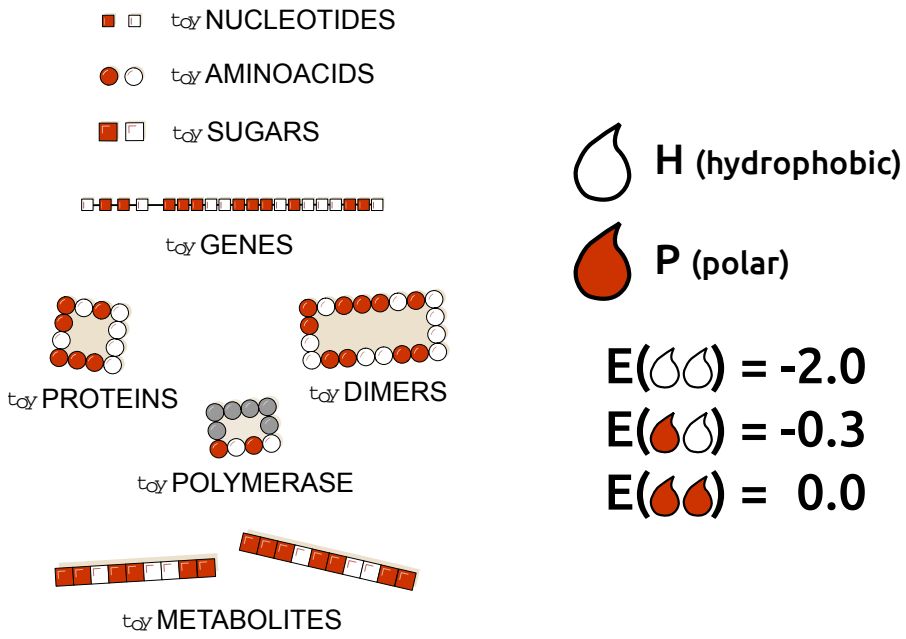


Figure 2.2: Building blocks and interactions defining toyLIFE. The three basic building blocks of toyLIFE are toyNucleotides, toyAminoacids, and toySugars. They can be hydrophobic (H, white) or polar (P, red), and their random polymers constitute toyGenes, toyProteins, and toyMetabolites. The toyPolymerase is a special polymer that will have specific regulatory functions. These polymers will interact between each other following an extension of the HP model (see text), for which we have chosen the interaction energies $E_{\mathrm{HH}} = -2$, $E_{\mathrm{HP}} = -0.3$ and $E_{\mathrm{PP}} = 0$ (Li et al., 1996).

2.1 Building blocks: genes, proteins, metabolites

The basic building blocks of toyLIFE are toyNucleotides (toyN), toyAminoacids (toyA), and toySugars (toyS). Each block comes in two flavors: hydrophobic (H) or polar (P). Random polymers of basic blocks constitute toyGenes (formed by 20 toyN units), toyProteins (chains of 16 toyA units), and toyMetabolites (sequences of toyS units of arbitrary length). These elements of toyLIFE are defined on two-dimensional space (Figure 2.2).

toyGenes

toyGenes are composed of a 4-toyN promoter region followed by a 16-toyN coding region. There are 2^4 different promoters and 2^{16} coding regions, leading to $2^{20} \approx 10^6$ toyGenes. An ensemble of toyGenes forms a genotype. If the toyGene is expressed, it will produce a chain of 16 toyA that represents a toyProtein. Translation follows a straightforward rule: H (P) toyN translate into H (P) toyA.

toyProteins

toyProteins correspond to the minimum energy, maximally compact folded structure of the 16 toyA chain arising from a translated toyGene. Their folded configuration is calculated through the hydrophobic-polar (HP) protein lattice model (Dill, 1985; Li et al., 1996).

We only consider maximally compact structures. That is, every toyProtein must fold into a 4×4 lattice, following a self-avoiding walk (SAW) on it. After accounting for symmetries —rotations and reflections—, there are only 38 SAWs on that lattice (Figure 2.3).

The energy of a fold is the sum of all pairwise interaction energies between toyA that are not contiguous along the sequence. Pairwise interaction energies are $E_{\mathrm{HH}} = -2$, $E_{\mathrm{HP}} = -0.3$ and $E_{\mathrm{PP}} = 0$, following the conditions set in Li et al. (1996) that $E_{\mathrm{PP}} > E_{\mathrm{HP}} > E_{\mathrm{HH}}$ (Figure 2.3). toyProteins are identified by their folding energy and their perimeter. If there is more than one fold with the same minimum energy, we select the one with fewer H toyAminoacids in the perimeter. If still there is more than one fold fulfilling both conditions, we discard that protein by assuming that it is intrinsically disordered and thus non-functional (Radivojac et al., 2007). Note, however, that sometimes different folds yield the same folding energy and the same perimeter. In those cases, we do not discard the resulting toyProtein [1]. Out of $2^{16} = 65,536$ possible toyProteins, $12,987$ do not yield unique folds. We find $2,710$ different toyProteins with 379 different perimeters. Not all toyProteins are equally abundant: although

[1] In Arias et al. (2014), where we first presented toyLIFE, we did not use this rule: whenever a sequence folded into two folds with the same folding energy and same number of Hs in the perimeter, we would discard them. This version of toyLIFE, therefore, is slightly different. However, the results are qualitatively similar.

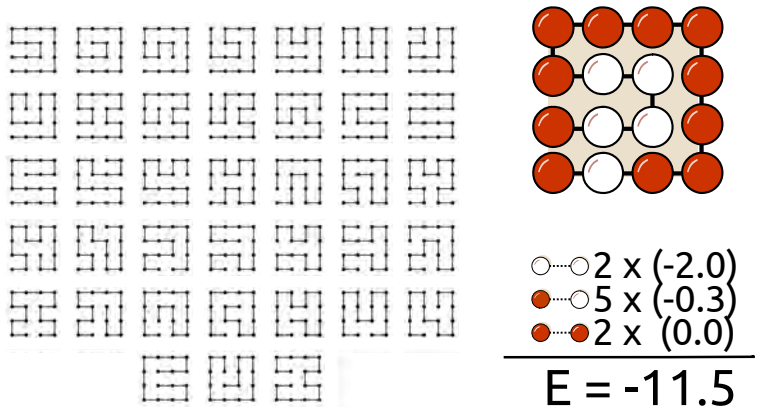


Figure 2.3: Protein folding in toyLIFE. toyProteins fold into a 4×4 lattice, following a self-avoiding walk (SAW). After accounting for symmetries, there are 38 SAWs (left). For each binary sequence of length 16, we fold it into every SAW and compute its folding energy, following the HP model. For instance, we fold the sequence PHPPPPPPPPPHHHHP into one of the SAWs and compute its folding energy (right). There are two HH contacts, five HP contacts and two PP contacts —we only take into account contacts between non-adjacent toyAminoacids. Summing all this contacts with their corresponding energies, we obtain a folding energy of -11.5. Repeating this process for every SAW, we obtain the minimum free structure.

every toyProtein is coded by 19.40 toyGenes on average, most of them are coded by only a few toyGenes. For instance, $1,364$ toyProteins —roughly half of them!— are coded by less than 10 toyGenes. On the other hand, only 4 toyProteins are coded by more than 200 toyGenes, the maximum being 235 toyGenes coding for the same toyProtein. The distribution is close to an exponential decay (Figure 2.4**a**). The same happens with the perimeters, although with less skewness: each perimeter is mapped by 7.15 toyProteins on average, but the most abundant perimeters correspond to 26 toyProteins, and 100 are mapped by 1 or 2 toyProteins (Figure 2.4**b**). As we will see later, this already induces a certain degree of neutrality in toyLIFE phenotypes.

Folding energies range from -18.0 to -0.6, with an average in -9.63. The distribution is unimodal, although very rugged (Figure 2.4**c**). Note that folding energies are discrete, and that separations between them are

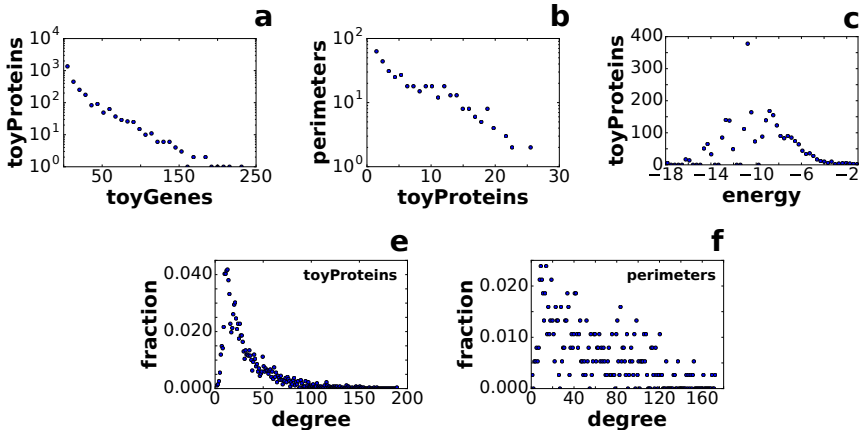


Figure 2.4: Distributions of toyProteins in toyLIFE. (a) Distribution of toyProtein abundances —that is, the number of toyGenes that code for them. Most toyProteins are coded by few toyGenes, but some of them are very abundant: the most abundant toyProtein is coded by 235 toyGenes. **(b)** Distribution of the perimeters associated with each toyProtein. Again, not all perimeters are equally abundant, and some of them correspond to as many as 25 toyProteins, while 100 correspond to 1 or 2 toyProteins. **(c)** Distribution of folding energies. The range of folding energies goes from -18.0 to -0.6, with a unimodal, rugged distribution. The mode is -10.6, a folding energy achieved by 202 toyProteins. **(d)** Degree distribution in the toyProtein network. Two toyProteins are connected if there are two toyGenes coding for them that have the same sequence, except for one toyN. The average degree is 32.2. **(e)** Degree distribution in the perimeter network. Two perimeters are neighbors if the toyProteins associated to them are neighbors. The average degree is 53.3.

not equal. For instance, there are 6 toyProteins that have a folding energy of -18.0, but the next energy level is -16.3, realized by 17 toyProteins, and yet the next level is -16.0, realized by 14 toyProteins. The mode of the distribution is -10.6, realized by 202 toyProteins.

We can also study the structure of the toyProtein network (Figure 2.4**e**, **f**). The nodes of this network will be the $2,710$ toyProteins. toyProtein 1 and toyProtein 2 will be neighbors if there is a pair of toyGenes that express each toyProtein and whose sequence is equal but for one toyN. The weight of the edge between toyProtein1 and 2 will be the sum of such pairs of toyGenes. Building this network, it is surprising that there are no auto-

loops —there are no mutations connecting one toyProtein to itself. In other words, although there is a strong degeneracy in the mapping from toyGenes to toyProteins, there are no connected neutral networks. If we consider just the perimeters, however, the neutrality is somewhat recovered: out of the 379 perimeters, 224 of them have neutral neighbors. So there are many mutations that alter the folding energy of a toyProtein without changing the perimeter. In this sense, tOYLIFE is capturing a complex detail of molecular biology: mutations can seem neutral from one point of view —in this case, perimeter— but are rarely entirely neutral. There are always some small changes in the molecule —in this case, folding energy—, that may affect their function later on. Real world examples of this *cryptic* effects of mutations on molecules are everywhere (Aharoni et al., 2005; Amitai et al., 2007; Khersonsky and Tawfik, 2010; Hayden et al., 2011). Connections between toyProteins are scarce, too: the average degree in the toyProtein network is 32.2 (with a standard deviation of 25.7), a very small number — on average, each toyProtein is connected to hardly 1% of the rest of toyProteins! (Figure 2.4**e**). The maximum degree is 190. This means that mutating from one toyProtein to other is not easy in general. In terms of perimeters this is more relaxed, as the average degree in the perimeter network is 53.3 (standard deviation is 38.1), with a maximum degree of 173. On average, every perimeter is connected to 14% of the rest of perimeters: it is a small number, but it is still higher than in the toyProtein case (Figure 2.4**f**).

In the tOYLIFE universe, only the folding energy and perimeter of a toyProtein matter to characterize its interactions, so folded chains sharing these two features are indistinguishable. This is a difference with respect to the original HP model, where different inner cores defined different proteins and the composition of the perimeter was not considered as a phenotypic feature. However, subsequent versions of HP had already included additional traits (Hoque et al., 2009).

The toyPolymerase (Figure 2.2) is a special toyA polymer, similar to a toyProtein in many aspects, but that is not coded for by any toyGene. It has only one side, with sequence PHPH, and its folding energy is -11.0. We will discuss its function and place later on.

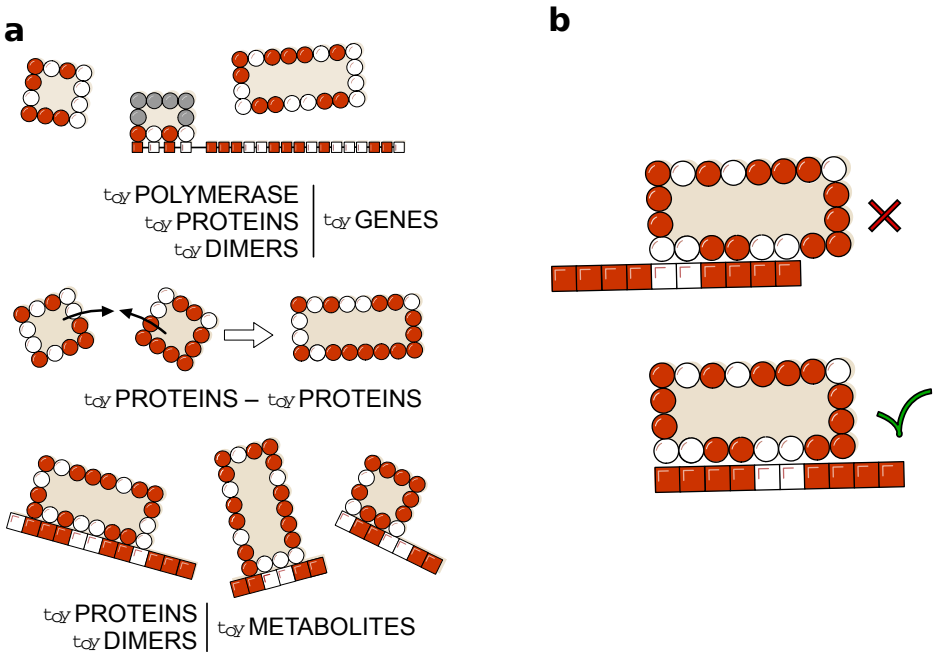


Figure 2.5: Interactions in toyLIFE. (a) Possible interactions between pairs of toyLIFE elements. toyGenes interact through their promoter region with toyProteins (including the toyPolymerase and toyDimers); toyProteins can bind to form toyDimers, and interact with the toyPolymerase when bound to a promoter; both toyProteins and toyDimers can bind a toyMetabolite at arbitrary regions along its sequence. **(b)** When a toyDimer or toyProtein binds to a toyMetabolite with the same energy in many places, we choose the most centered binding. If both positions are equally centered, then no binding occurs.

2.2 Extending the HP model: interactions

toyProteins interact through any of their sides with other toyProteins, with promoters of toyGenes, and with toyMetabolites (see Figure 2.5**a**). When toyProteins bind to each other, they form a toyDimer, which is the only protein aggregate considered in toyLIFE. The two toyProteins disappear, leaving only the toyDimer. Once formed, toyDimers can also bind to promoters or toyMetabolites through any of their sides —binding to other toyProteins or toyDimers, however, is not permitted. In all cases, the interaction energy (E_{int}) is the sum of pairwise interactions for all HH, HP and PP pairs

formed in the contact —these interactions follow the rules of the HP model as well. Bonds can be created only if the interaction energy between the two molecules E_{int} is lower than a threshold energy $E_{\text{thr}} = -2.6$. Note that a minimum binding energy threshold is necessary to avoid the systematic interaction of any two molecules. Low values of the threshold would lead to many possible interactions, which would increase computation times. High values would lead to very few interactions, and we would obtain a very dull model. Our choice of $E_{\text{thr}} = -2.6$ achieves a balance: the number of interactions is large enough to generate complex behaviors, as we will see later on, while at the same time keeping the universe of interactions small enough to handle computationally. Instead of adding a threshold, we could have added terms that represent an energetic cost (in other models, as in RNA folding, the threshold is set to 0 because structural elements such as loops or dangling ends yield positive contributions to the total folding energy) or considered stochastic interactions, such that those with higher energy would be less probable. If below threshold, the total energy of the resulting complex is the sum of E_{int} plus the folding energy of all toyProteins involved. The lower the total energy, the more stable the complex. When several toyProteins or toyDimers can bind to the same molecule, only the most stable complex is formed. Consistently with the assumptions for protein folding, when this rule does not determine univocally the result, no binding is produced.

As the length of toyMetabolites is usually longer than 4 toyS (the length of interacting toyProtein sites), several binding positions between a toyMetabolite and a toyProtein might share the same energy. In those cases we select the sites that yield the most centered interaction (Figure 2.5**b**). If ambiguity persists, no bond is formed. Also, no more than one toyProtein / toyDimer is allowed to bind to the same toyMetabolite, even if its length would permit it. toyProteins / toyDimers bound to toyMetabolites cannot bind to promoters.

Interaction rules in tOYLIFE have been devised to remove any ambiguity. When more than one rule could be chosen, we opted for computational simplicity, having made sure that the general properties of the model remained unchanged. A detailed list of the specific disambiguation rules implemented in the model follows:

1. **Folding rule:** if a sequence of toyAminoacids can fold into two (or more) different configurations with the same energy and two different perimeters with the same number of H, it is considered degenerate and does not fold.

2. **One-side rule:** any interaction in which a toyProtein can bind any ligand with two (or more) different sides and the same energy is discarded.

3. **Annihilation rule:** if two (or more) toyProteins can bind a ligand with the same energy, the binding does not occur. However, if a third toyProtein can bind the ligand with greater (less stable) energy than the other two, and does so uniquely, it will bind it.

4. **Identity rule:** an exception to the Annihilation rule occurs if the competing toyProteins are the same. In this case, one of them binds the ligand and the other(s) remains free.

5. **Stoichiometric rule:** an extension of the Identity rule. If two (or more) copies of the same toyProtein / toyDimer / toyMetabolite are competing for two (or more) different ligands, there will be binding if the number of copies of the toyProtein / toyDimer / toyMetabolite equals the number of ligands. For example, say that P1 binds to P2, P3 and P4 with the same energy. Then, (a) if P1, P2 and P3 are present, no complex will form; (b) if there are two copies of P1, dimers P1-P2 and P1-P3 will both form; but (c) if P4 is added, no complex will form. Conversely, if all ligands are copies as well, the Stoichiometry rule does not apply. For example, three copies of P1 and two copies of P2 will form two copies of dimer P1-P2, and one copy of P1 will remain free.

2.3 Regulation

Expression of toyGenes occurs through the interaction with the toyPolymerase, which is a special kind of toyProtein (see Figure 2.2). The toyPolymerase only has one interacting side (with sequence PHPH) and its folding energy is fixed to value -11.0: it is more stable than more than half the

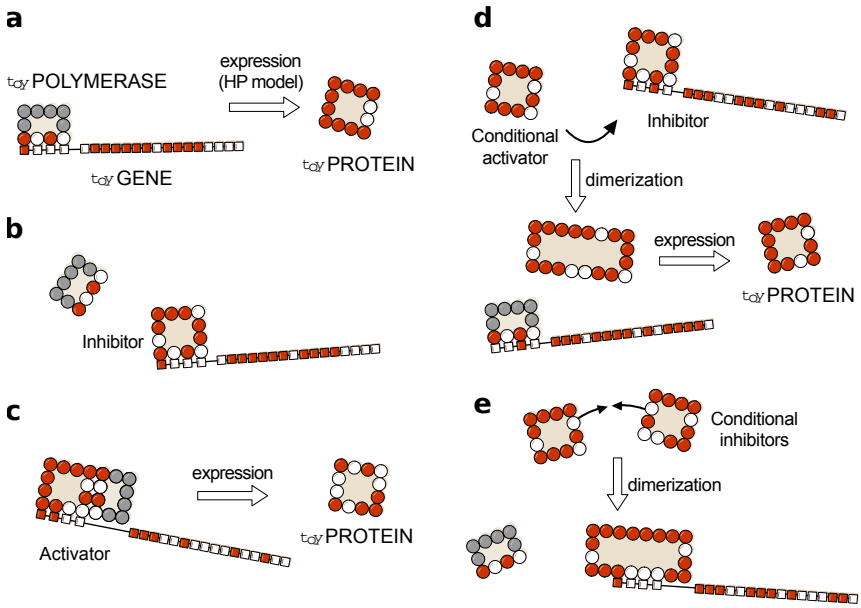


Figure 2.6: Regulatory functions in toyLIFE. (a) A toyGene is expressed (translated) when the toyPolymerase binds to its promoter region. The sequence of Ps and Hs of the toyProtein will be exactly the same as that of the toyGene coding region. **(b)** If a toyProtein binds to the promoter region of a toyGene with a lower energy than the toyPolymerase does, it will displace the latter, and the toyGene will not be expressed. This toyProtein acts as an *inhibitor*. **(c)** The toyPolymerase does not bind to every promoter region. Thus, not all toyGenes are expressed constitutively. However, some toyProteins will be able to bind to these promoter regions. If, once bound to the promoter, they bind to the toyPolymerase with their rightmost side, the toyGene will be expressed, and these toyProteins act as *activators*. **(d)** More complex interactions —involving more elements— appear. For example, a toyProtein that forms a toyDimer with an inhibitor —preventing it from binding to the promoter— will effectively activate the expression of the toyGene. However, it does neither interact with the promoter region nor with the toyPolymerase, and its function is carried out only when the inhibitor is present. We call this kind of toyProteins *conditional activators*. **(e)** Two toyProteins can bind together to form a toyDimer that inhibits the expression of a certain toyGene. As they need each other to perform this function, we call them *conditional inhibitors*. As the number of genes increases, this kind of complex relationships can become very intricate.

toyProteins. It is always present in the system. The toyPolymerase binds to promoters or to the right side of a toyProtein / toyDimer already bound to a promoter. When the toyPolymerase binds to a promoter, translation is directly activated and the corresponding toyGene is expressed (Figure 2.6**a**). However, a more stable (lower energy) binding of a toyProtein or toyDimer to a promoter precludes the binding of the toyPolymerase. This inhibits the expression of the toyGene, except if the toyPolymerase binds to the right side of the toyProtein / toyDimer, in which case the toyGene can be expressed.

The minimal interaction rules that define toyLIFE dynamics endow toyProteins with a set of possible activities not included *a priori* in the rules of the model (see Figure 2.6). For example, since the 4-toyN interacting site of the toyPolymerase cannot bind to all promoter regions —because some of these interactions have $E_{\text{int}} > E_{\text{thr}}$—, translation mediated by a toyProtein or toyDimer binding might allow the expression of genes that would otherwise never be translated. These toyProteins thus act as activators (Figure 2.6**c**). This process finds a counterpart in toyProteins that bind to promoter regions more stably than the toyPolymerase does, and therefore prevent gene expression —this happens if $E_{\text{int(PROT)}} + E_{\text{PROT}} < E_{\text{int(POLY)}} + E_{\text{POLY}}$. They are acting as inhibitors (Figure 2.6**b**). There are two additional functions that could not be foreseen and involve a larger number of molecules. A toyProtein that forms a toyDimer with an inhibitor —preventing its binding to the promoter— effectively behaves as an activator for the expression of the toyGene. However, it interacts neither with the promoter region nor with the toyPolymerase, and its activating function only shows up when the inhibitor is present. This toyProtein thus acts as a conditional activator (Figure 2.6**d**). On the other hand, two toyProteins can bind together to form a toyDimer that inhibits the expression of a particular toyGene. As the presence of both toyProteins is needed to perform this function, they behave as conditional inhibitors (Figure 2.6**e**). This flexible, context-dependent behavior of toyProteins is reminiscent of phenomena observed in real cells (Piatigorsky, 2007), and permits the construction of complex toyGene Regulatory Networks (toyGRNs).

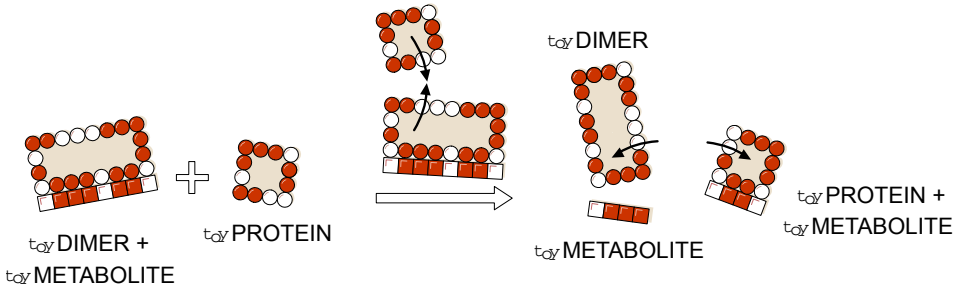


Figure 2.7: Metabolism in toyLIFE. A toyDimer is bound to a toyMetabolite when a new toyProtein comes in. If the new toyProtein binds to one of the two units of the toyDimer, forming a new toyDimer energetically more stable than the old one, the two toyProteins will unbind and break the toyMetabolite up into two pieces. We say that the toyMetabolite has been catabolized.

2.4 Metabolism

A first version of toyLIFE stopped at the regulatory level. The original aim of the model was to connect low-level mutational dynamics to high-level regulatory phenotypes. Once the model was designed, however, we decided to include a simple, primitive metabolism, in order to gain insight into the relationship between high-level phenotypes and fitness. As a consequence, the full definition of toyLIFE includes a primitive catabolism, carried out by toyDimers in conjunction with toyProteins. When a toyDimer is bound to a toyMetabolite, another toyProtein can interact with this complex and break it. This reaction will take place if the toyProtein can bind to one of the subunits of the toyDimer and the resulting complex has less total energy than the toyDimer. As with the rest of interactions, the catabolic reaction will only take place if this binding is unambiguous. As a result of this reaction, the toyDimer will be broken in two: one of the pieces will be bound to the toyProtein, and the other one will remain free. The toyMetabolite will break accordingly: the part of it that was bound to the first subunit will stay with it, and the other part will stay with the second subunit. Note that the toyMetabolite need not be broken symmetrically: this will depend on how the toyDimer binds to it (Figure 2.7).

This definition of metabolism opens the door to a relationship with the environment in toyLIFE, mediated by toyMetabolites. We will briefly

explore this relationship in following chapters, although the original aim of studying the relationship between phenotypes and fitness has proven too large for the present thesis —see Chapter 7, however, for some ideas on how to develop these ideas in the future.

2.5 Dynamics in toyLIFE

The dynamics of the model proceeds in discrete time steps and variable molecular concentrations are not taken into account. A step-by-step description of toyLIFE dynamics is summarized in Figure 2.8. There is an initial set of molecules which results from the previous time step: toyProteins (including toyDimers and the toyPolymerase) and toyMetabolites, either endogenous or provided by the environment. These molecules first interact between them to form possible complexes (see Section 2.2) and are then presented to a collection of toyGenes that is kept constant along subsequent iterations. Regulation takes place, mediated by a competition for binding the promoters of toyGenes, possibly causing their activation and leading to the formation of new toyProteins. Binding to promoters is decided in sequence. Starting with any of them (the order is irrelevant), it is checked whether any of the toyProteins / toyDimers (including the toyPolymerase) available bind to the promoter —remember that complexes bound to toyMetabolites are not available for regulation—, and then whether the toyPolymerase can subsequently bind to the complex and express the accompanying coding region. If it does, the toyGene is marked as active and the toyProtein / toyDimer is released. Then a second promoter is chosen and the process repeated, until all promoters have been evaluated. toyGenes are only expressed after all of them have been marked as either active or inactive. Each expressed toyGene produces one single toyProtein molecule. There can be more units of the same toyProtein, but only if multiple copies of the same toyGene are present.

toyProteins / toyDimers not bound to any toyMetabolite are eliminated in this phase. Thus, only the newly expressed toyProteins and the complexes involving toyMetabolites in the input set remain. All these molecules interact yet again, and here is where catabolism can occur. Catabolism happens when, once a toyMetabolite-toyDimer complex is formed, an additional toyProtein binds to one of the units of the toyDimer with an energy

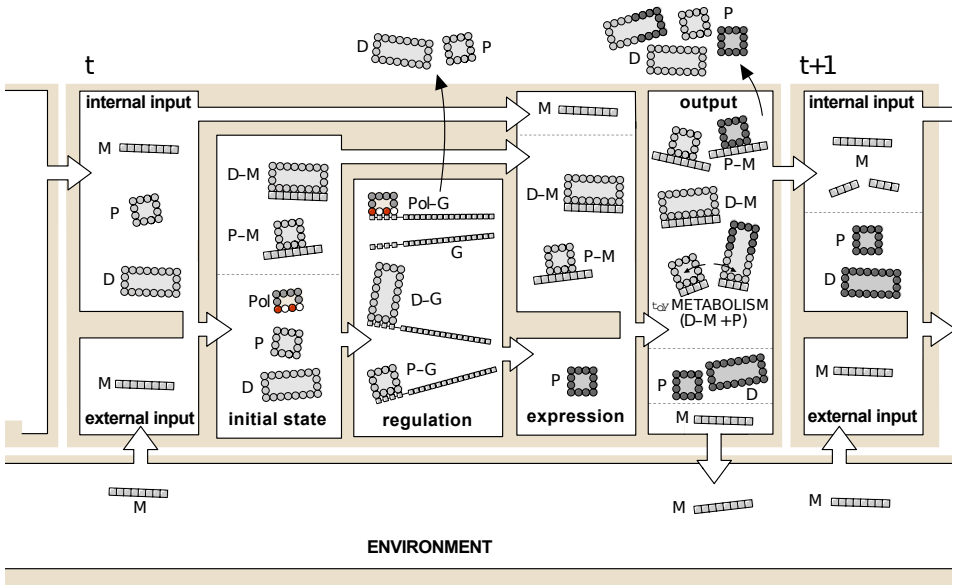


Figure 2.8: Dynamics of toyLIFE. Input molecules at time step t are toyProteins (Ps) (including toyDimers (Ds)) and toyMetabolites, either produced as output at time step $t-1$ or environmentally supplied (all toyMetabolites denoted Ms). Ps and Ds interact with Ms to produce complexes P-M and D-M. Next, the remaining Ps and Ds and the toyPolymerase (Pol) interact with toyGenes (G) at the regulation phase. The most stable complexes with promoters are formed (Pol-G, P-G and D-G), activating or inhibiting toyGenes. P-Ms and D-Ms do not participate in regulation. Ps and Ds not in complexes are eliminated and new Ps (dark grey) are formed. These Ps interact with all molecules present and form Ds, new P-M and D-M complexes, and catabolize old D-M complexes. At the end of this phase, all Ms not bound to Ps or Ds are returned to the environment, and all Ps and Ds in P-M and D-M complexes unbind and are degraded. The remaining molecules (Ms just released from complexes, as well as all free Ps and Ds) go to the input set of time step $t+1$.

that is lower than that of the initial toyDimer. In this case, the latter disassembles in favor of the new toyDimer, and in the process the toyMetabolite is broken, as already mentioned in Section 2.4 and Figure 2.7. The two pieces of the broken toyMetabolites will contribute to the input set at the next time step, as will free toyProteins / toyDimers. However, toyProteins / toyDimers bound to toyMetabolites disappear in this phase —they are

degraded—, and only the toyMetabolites are kept as input to the next time step. Unbound toyMetabolites are returned to the environment. This way, the interaction with the environment happens twice in each time step: at the beginning and at the end of the cycle.

2.6 GRNs in toyLIFE are deterministic Boolean networks

Molecular interactions and dynamical rules in toyLIFE can be translated into toyGRNs that behave as deterministic Boolean networks (Kauffman, 1969; Cheng et al., 2011). The corresponding Boolean variables are the states (expressed or not expressed) of toyGenes. These variables are transformed through Boolean functions that represent the dynamical rules described in the previous section, having as input current toyGene states and as output their states at the next time step. Boolean functions depend on the toyProteins present in the system and on the functions they perform. Through iteration of the Boolean map one can characterize the set of attractors of the dynamics and the corresponding basins of attraction.

If the initial set is formed by k genes, we should consider 2^k different possible vectors of dimension k that correspond to the initial states (i.e. all combinations of genes being expressed (1) or not expressed (0)). First, the presence of possible toyDimers coming from expressed genes is evaluated, and then their interactions with promoter regions (in competition or cooperation with the toyPolymerase and other toyProteins) are evaluated. This yields an updated set of expressed toyGenes (a different state) to which the previous rules are again applied. In this way, one can construct a truth table that can be subsequently represented in the form of a directed graph (indicating which state maps into which other) and is fully analogous to a deterministic Boolean network. An example of a Boolean network derived from a system of three genes is represented in Figure 2.9.

The presence of toyMetabolites may modify toyGRNs by changing the output states of the corresponding Boolean network (Figure 2.10). According to the dynamical rules of toyLIFE, toyMetabolites may interact with toyProteins or toyDimers. Any molecule bound to a toyMetabolite is no longer available to bind to promoters, and therefore the expression of the toyGRN is modified. An example of how a toyGRN might change can be derived from Figure 2.9: if a toyMetabolite able to bind to toyDimer 1-3

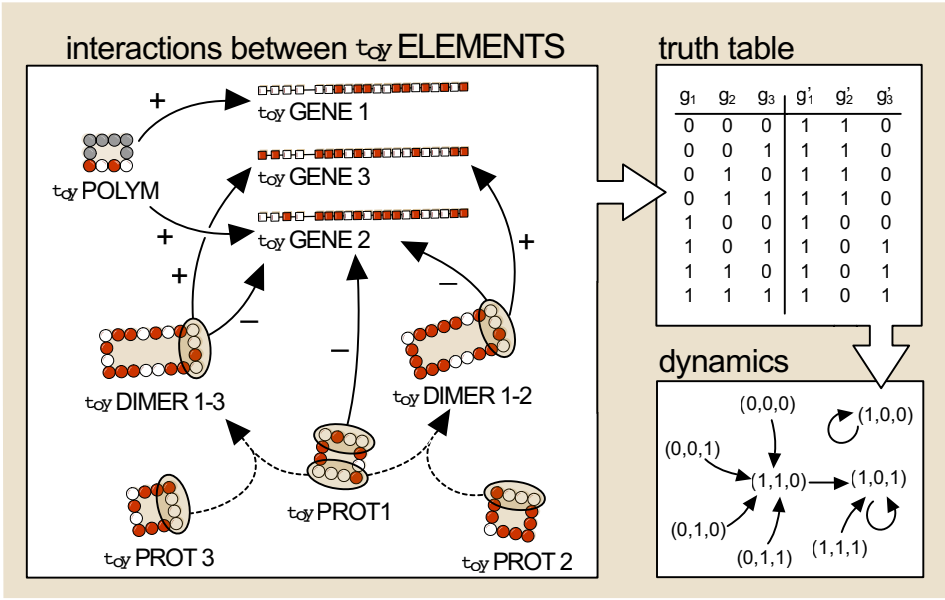


Figure 2.9: Example of a Boolean network produced by toyLIFE rules. The inputs of the truth table (possible initial states) are all combinations of states of three toyGenes. Whenever a toyGene is active, the toyProtein it codes for is present. The main panel schematically represents all relevant interactions between molecules: in this case the toyPolymerase may bind to the promoter regions of toyGenes 1 and 2 ($+$ signs), and toyProtein 1 inhibits the expression of toyGene 2 ($-$ signs). The simultaneous presence of toyProteins 1 and 3 leads to toyDimer 1-3, and the simultaneous presence of toyProteins 1 and 2 to toyDimer 1-2. Both toyDimers inhibit the expression of toyGene 2 and activate the expression of toyGene 3. The construction of the Boolean functions codified in the truth table is straightforward given the interactions conditional on presence or absence of each toyProtein. The truth table maps every possible initial state (g_i) to its corresponding regulatory output (g_i'). When the truth table is represented as a directed graph (summarizing the dynamics of the system from all possible initial conditions) it is seen that there are two attractors for the dynamics: $(1,0,1)$, whose basin of attraction has size 7, and $(1,0,0)$, whose basin of attraction has size 1. (Note that the order of toyGenes in a genome is irrelevant, and only responds to aesthetic reasons.)

is added to the input set, state $(1,0,1)$ is mapped to $(1,1,0)$ (Figure 2.10), and not to itself as before.

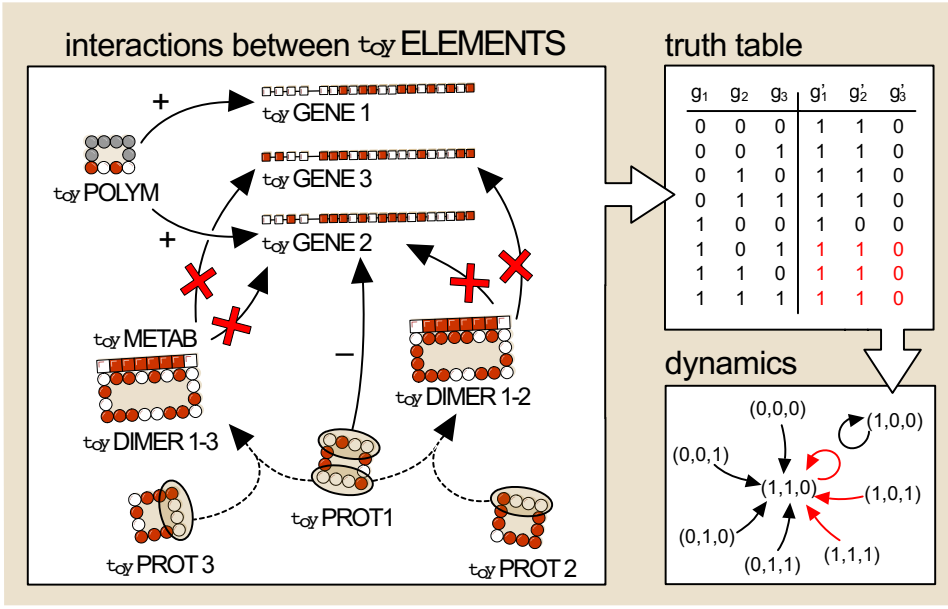


Figure 2.10: toyMetabolites change the expression of toyGRNs. This is the same example illustrated in Figure 2.9, but with the addition of a toyMetabolite able to bind toyDimers 1-2 and 1-3. When these toyDimers bind to the toyMetabolite, they no longer participate in the regulation phase, and thus states $(1,0,1)$, $(1,1,0)$ and $(1,1,1)$ are all mapped to state $(1,1,0)$ in the presence of this toyMetabolite. In other words, the presence of the toyMetabolite changes three entries in the truth table, and therefore the associated Boolean network —whose asymptotic state is now a different one.

2.7 Example

Let us explore how all the elements presented in this chapter come together to give a coherent model. We will use the same genotype as in Figures 2.9 and 2.10. A schematic of the process is shown in Figure 2.11.

For convenience, we assume the initial state to be that in which all toyGenes are off. Of course, different initial states can be taken into account, generating different dynamics. As a result, the input set at time step t_0 is empty. When the regulation phase arrives, only the toyPolymerase is present to bind to the promoters. In this example, it can bind to the promoters of toyGenes 1 and 2, activating their expression. Therefore, toyPro-

teins 1 and 2 are expressed and, in the next phase, they form toyDimer 1-2. Assuming no other input from the environment, the only molecule in the output set is toyDimer 1-2.

At time step t_1, the internal input set only contains toyDimer 1-2. No other molecule comes in from the environment, and we go into the regulation phase. The toyPolymerase activates the expression of toyGene1, but toyDimer 1-2 inhibits the expression of toyGene 2, and also activates that of toyGene 3. As a result, toyDimer 1-3 is formed, and it will be the only molecule in the output set.

At time step t_2, the internal input set only contains toyDimer 1-3, and we keep assuming no other molecules come in from the environment. In the regulation phase, toyDimer 1-3 has the same role as toyDimer 1-2: toyProtein 1 and 3 are expressed, and toyDimer 1-3 is formed again. It is clear that, if there is no change in the environment, toyDimer 1-3 will be expressed time step after time step, *ad infinitum*. The toyGRN has reached a steady state.

Now, let us explore what happens when a toyMetabolite comes in from the environment. We will assume that the environment changes, and that a constant dose of one toyMetabolite will enter into the cell every time step. The toyGRN will start in the steady state already described. At time step t'_0, the internal input set consists of toyDimer 1-3, and the external input set consists of one copy of the toyMetabolite. This toyMetabolite is such that it binds toyDimer 1-3, which is therefore unable to participate in the regulation phase. Without any competition, the toyPolymerase activates toyGenes 1 and 2, as before. toyProtein 1 is able to bind toyDimer 1-3, breaking it in two. As we mentioned in Section 2.5, the remnants of toyDimer 1-3 and toyProtein 1 will be discarded at the end of this time step.

The internal input at time step t'_1 will consist of toyProtein 2 and the rests of the toyMetabolite. The external input set, again, contains one molecule of toyMetabolite. These molecules interact with each other, but toyProtein 2 cannot bind any of the toyMetabolites, so it goes directly into the regulation phase. toyProtein 2 has no effect on regulation, and again toyProteins 1 and 2 are expressed, and toyDimer 1-2 is formed. Note that toyDimer 1-2 cannot interact with the toyMetabolites in this phase, because it has just been formed. In other words, the interaction phase consists of

toyProteins 1 and 2 and all the toyMetabolites. toyProteins 1 and 2 prefer to form the toyDimer instead of binding to the toyMetabolites: the output of the interaction phase is the toyDimer 1-2, and there is no subsequent interaction until the next time step.

As no molecule has bound the toyMetabolites, they will not be present in the internal input set of time step t_2', which will only contain toyDimer 1-2. If a new toyMetabolite is provided as the external input, the toyDimer will bind to it, and the cycle begins again (however, note that from now on all metabolism will be due to toyDimer 1-2 instead of toyDimer 1-3). This genotype will be able to metabolize the toyMetabolite as long as it is present in the environment.

2.8 Definition of phenotype

What is the phenotype in toyLIFE? This is a relevant question, given that we are defining toyLIFE as a model of the genotype-phenotype map.

We already discussed (in Chapter 1) the difficulties associated to the definition of phenotype in the era of cell biology. It is evident that a comprehensive definition, including all aspects of cellular biology, is not useful for most applications. In the case of toyLIFE, it will be difficult to extract information from a complex definition of the phenotype, involving toyProteins, interactions, truth tables and metabolic abilities. Therefore, we need to simplify things.

In this thesis, we will use two different definitions of phenotype for toyLIFE. The first one is metabolic, and refers to the set of toyMetabolites that a given genotype can catabolize, after it has reached the regulatory equilibrium, starting from the initial state when all toyGenes are off —that is, the same conditions presented in the previous section. The space of toyMetabolites is infinite, as their length is arbitrary. However, because the longest interacting side of a toyDimer is 8 toyA long, toyMetabolites longer than 8 toyS will include inside themselves a subsequence equivalent to a toyMetabolite of length 8 or smaller. Therefore, we will only study the space of toyMetabolites up to length 8. The first definition will be the focus of Chapters 3 and 4.

The second definition is regulatory. We will consider a spatial arrangement of toy-cells —cells containing toyGenes and functioning under toy-

LIFE rules— in one dimension. Every one of these toy-cells will share the same genotype. The difference with respect to isolated cells is that each cell in the array receives its input molecules from their adjacent neighbors. The phenotype will be the spatio-temporal pattern of expression they generate after a given input. In Chapter 5 we will see how this definition is closely related to cellular automata.

Both definitions are arbitrary, but are sufficiently interesting to explore many properties of the genotype-phenotype map in $\mathtt{toy}$LIFE.

2.9 Summary

In this chapter, we have presented $\mathtt{toy}$LIFE, a multi-level model for the genotype-phenotype map. $\mathtt{toy}$LIFE contains genes, proteins and metabolites that interact through the laws of a simplified chemistry, forming complex regulatory and metabolic networks. It is intended to bridge the gap between low-level models of the genotype-phenotype map, such as RNA or protein secondary structure, with high-level models, such as regulatory and metabolic networks.

$\mathtt{toy}$LIFE is a complex model, even in its simplicity. It extends the rules of the HP model to build a caricature of cell biology. In devising the model here, we had to make some choices regarding energy parameters, number of molecules or genes allowed to interact, or disambiguation rules to define functional molecules. We do not claim that $\mathtt{toy}$LIFE matches biological reality, and it was not our intention to do so. The interaction and dynamical rules in $\mathtt{toy}$LIFE have been chosen so as to make the model as simple as possible, while retaining the essential characteristics of molecular genetics. We aim to explore universal features of complex molecular systems, regardless of the details. In that sense, although similar models with different rules might be devised, we would expect that many of them (if not most) would display a phenomenology comparable to our $\mathtt{toy}$LIFE. The main principles behind the complex interactions between molecules, regulation and metabolism must be largely independent on these kind of details.

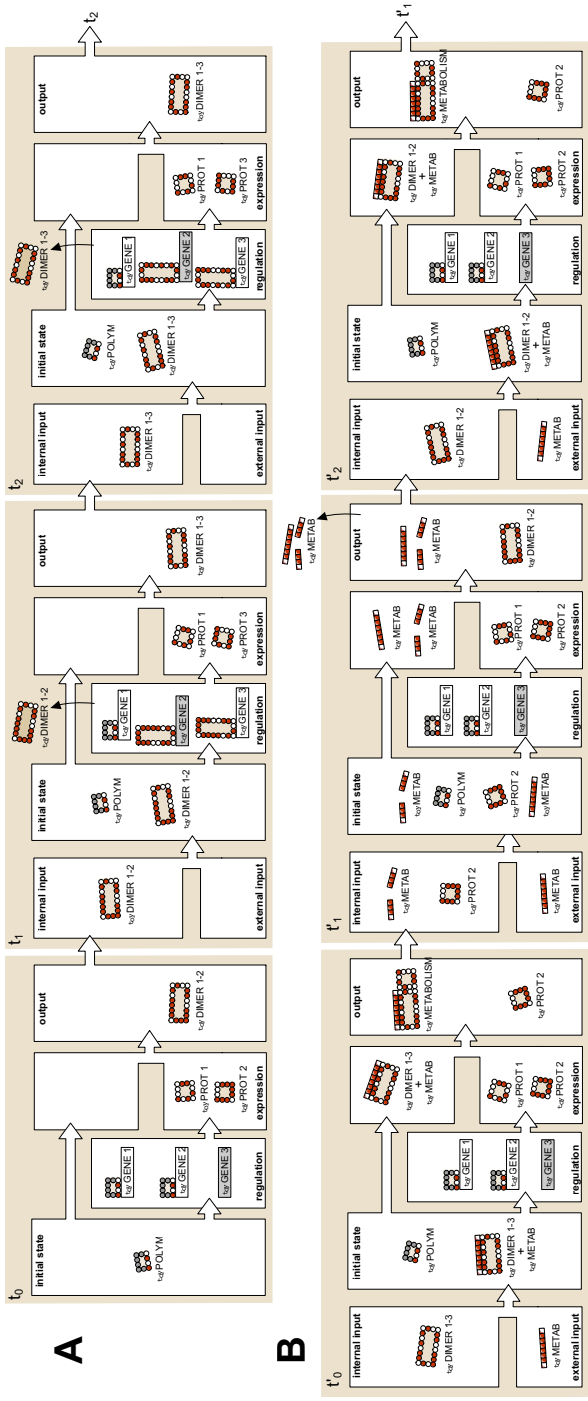

A
B
t0
t1
t2
t'0
t'1
t'2
initial state
internal input
external input
output
regulation
expression
tgf POLYM
tgf DIMER 1-2
tgf DIMER 1-3
tgf METAB
tgf PROT 1
tgf PROT 2
tgf PROT 3
tgf GENE 1
tgf GENE 2
tgf GENE 3
tgf METABOLISM

Figure 2.11: (Previous page.) **Summary of a metabolon activity in toyLIFE.** Consider the toyGRN of Figures 4 and 5 (**A**). Initially (t_0) all three toyGenes are off. The toyPolymerase can bind to the promoter regions of toyGenes 1 and 2, expressing toyProteins 1 and 2, and toyDimer 1-2 forms. Thus, the internal input set for time step t_1 contains toyDimer 1-2. At the regulation phase in t_1 the toyPolymerase (which is always present) activates the expression of toyGene 1, and toyDimer 1-2 inhibits the expression of toyGene 2 and activates that of toyGene 3. As a result, toyDimer 1-3 forms. The input set for time step t_2 then contains just toyDimer 1-3. At t_2 toyDimer 1-3 again inhibits the expression of toyGene 2 and activates that of toyGene 3, and the internal input set for next time step will again only contain toyDimer 1-3. The toyGRN has reached a steady state. But if at this point a toyMetabolite is added to the input set, the behavior of the toyGRN changes (**B**). The toyMetabolite is such that it binds toyDimer 1-3, so the toyDimer is unable to participate in regulation, and the toyPolymerase activates the expression of toyGenes 1 and 2. toyProtein 1 is then able to bind to toyDimer 1-3 in the output phase, breaking it. The internal input set for time step t_1' is formed by toyProtein 2 and the rests of the broken toyMetabolite. Even if the toyMetabolite appears again as a external output, no molecule can bind it in the input phase, so this does not affect regulation. toyProtein 2 has no effect on regulation, and again toyProteins 1 and 2 are expressed, and toyDimer 1-2 is formed. As no molecule has bound the toyMetabolites, they will not be present in the internal input set of time step t_2', which will only contain toyDimer 1-2. If a new toyMetabolite is provided as the external input, the toyDimer will bind to it, and the cycle begins again (however, note that from now on all metabolism will be due to toyDimer 1-2 instead of toyDimer 1-3).

3

The genotype-phenotype map in toyLIFE

"What a useful thing a pocket-map is!" I remarked.
"That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"
"About six inches to the mile."
"Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"
"Have you used it much?" I enquired.
"It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."

Lewis Carroll
Sylvie and Bruno Concluded (1895)

In this chapter we will explore the metabolic genotype-phenotype map in toyLIFE. We will use the metabolic definition of phenotype described at the end of Chapter 2: the set of metabolites from lengths 4 to 8 that a genotype can metabolize after reaching the regulatory equilibrium, starting from the state where all toyGenes are off.

The size of genotype space gets uncannily large with genotype size. For each number of toyGenes in the genotype g, a genotype is formed by choosing from the set of 2^{20} toyGenes with repetition. Because the order of the toyGenes is irrelevant, the number of genotypes we can form equals the number of combinations with repetition of g objects (toyGenes) in $k = 2^{20}$ categories, given by $\binom{g+2^{20}-1}{g}$. For $g = 2$, this number is 5.5×10^{11}. For $g = 3$, it is 1.9×10^{17}. For $g = 4$, it is 5.0×10^{22}, and for $g = 5$ it is 1.1×10^{28}. An exhaustive exploration of these genotype spaces is well over our computational possibilities. However, using computational tricks, we have exhaustively sampled the $g = 2$ and $g = 3$ cases. This is what we present now.

3.1 A note on toyMetabolites

There are 2^m binary strings —toyMetabolites— of length m. From lengths 4 to 8, therefore, there are

$$\sum_{m=4}^{8} 2^m = 496$$

toyMetabolites. However, due to the interaction rules of toyLIFE, a particular string and its reverse —i.e. HPPHPPPP and PPPPHPPH— will be treated the same way by toyLIFE organisms. Therefore, for all practical purposes, we will consider each string and its reverse as the same toyMetabolite, thus staying with 274 of them. Additionally, there are 60 toyMetabolites that cannot be catabolized in toyLIFE (Figure 3.1). For all lengths, toyMetabolites formed by all Ps and one H in the extrema, or all Hs and one P in the extrema, are unbreakable. This is because there is no unambiguous way in which a toyDimer can bind to these toyMetabolites. There are two of these toyMetabolites for each length, making a total

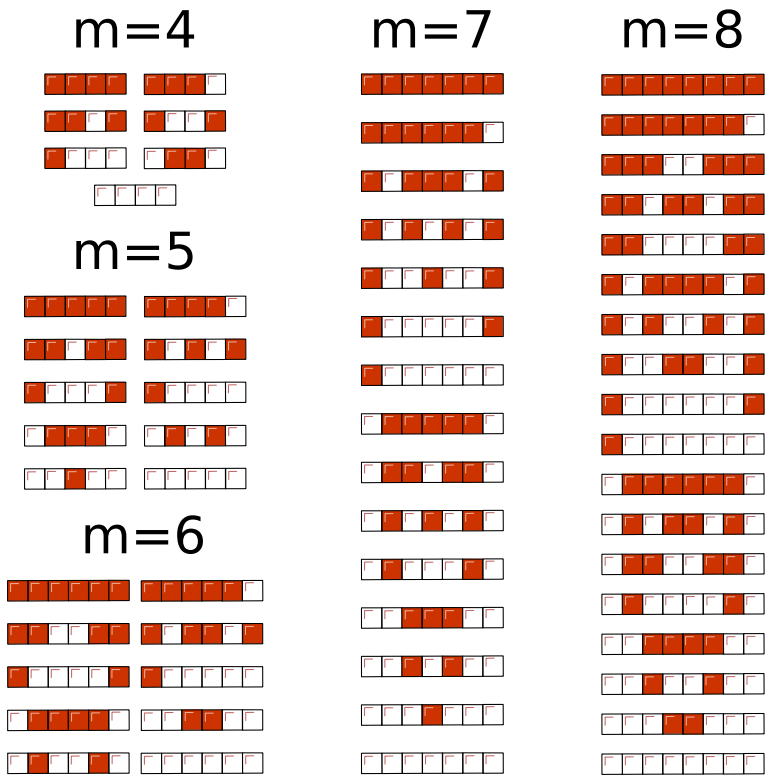


Figure 3.1: Unbreakable toyMetabolites. There are 60 unbreakable toyMetabolites: 49 of them are symmetrical, other 10 are chains of all Hs or all Ps in a row, and the last one is PPHP. Because of the interaction rules in toyLIFE, only symmetrical toyDimers will be able to bind these toyMetabolites, and therefore they cannot be broken.

of 10. Additionally, the toyMetabolite PPHP cannot be broken due to the same reason. Symmetrical toyMetabolites, in general, cannot be catabolized either. Because of the interaction rules described in Chapter 2, only symmetrical toyDimers can bind to these toyMetabolites. But symmetrical toyDimers cannot be broken: any toyProtein that can bind to one subunit will be able to bind the other one. Because of the disambiguation rules, no binding is produced, and catabolism does not occur. There are 52 symmet-

ric toyMetabolites —because they repeat half the sequence, there are

$$\sum_{m=4}^{8} 2^{\left[\frac{m+1}{2}\right]} = 52$$

of them, $[x]$ being the integer part of x —odd-length symmetrical toyMetabolites repeat $m+1$ toySugars, hence the $[(m+1)/2]$ exponent. However, three symmetrical toyMetabolites of length 7 —namely, PPPHPPP, PPHPHPP and PPHHHPP— can actually be broken. So there are 49 unbreakable symmetrical toyMetabolites. Added to the previous 11 unbreakable toyMetabolites, we get the total of 60. As a result, the total number of toyMetabolites up to length 8 is 214.

It is somewhat interesting that, as an emergent property of the model, some toyMetabolites are not able to be catabolized. Moreover, it is not that these toyMetabolites are irrelevant to the model: if they are present, they will interact with symmetric toyDimers, affecting the regulatory output of cells. So these toyMetabolites could function as signalling molecules.

3.2 Degeneracy of the genotype-phenotype map

For the $g = 2$ case, out of 5.5×10^{11} genotypes, only 1.1×10^{9} genotypes are able to catabolize any toyMetabolite, representing little more than 0.2% of all genotypes —one every 500. In the $g = 3$ case, out of 1.9×10^{17} genotypes, 1.0×10^{15} are able to break at least one toyMetabolite. This represents around 0.53% of all genotypes formed by $g = 3$ toyGenes — only one every 200. In both cases, the great majority of genotypes are unable to catabolize any toyMetabolite. But note that the space of *viable* genotypes is huge anyway.

Among these viable genotypes, there is an enormous degeneracy: many genotypes express the same metabolic phenotype. Thus, for $g = 2$, there are only 775 phenotypes, corresponding to an average of 1.4×10^{6} genotypes per phenotype. As for $g = 3$ there are 26,492 phenotypes, corresponding to an average of 3.8×10^{10} genotypes per phenotype, a huge degeneracy. From now on, we will refer to the set of phenotypes in $g = 2$ and $g = 3$ space as $\mathcal{P}_2$ and $\mathcal{P}_3$, respectively.

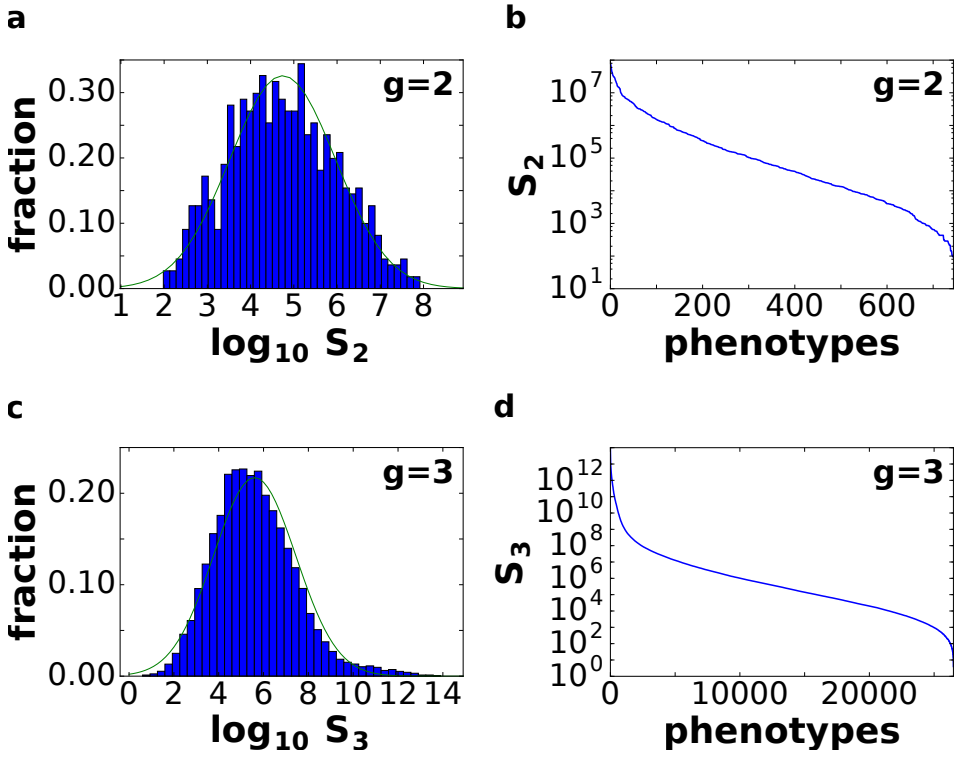


Figure 3.2: Degeneracy and asymmetry in the genotype-phenotype map in toyLIFE. (a) The distribution of sizes of phenotypes for $g = 2$ (S_2) follows a log-normal law, whose probability density function is: $f(x) = (x\sigma\sqrt{2\pi})^{-1}\exp(-(\log x - \mu)^2/2\sigma^2)$, where μ is the mean and σ is the standard deviation of the normally distributed logarithm of the variable. Here $\mu = 4.742$ and $\sigma = 1.224$ (empirically obtained from the log-transformed size distribution). **(b)** The rank distribution shows a long tail of rare phenotypes. **(c)** For $g = 3$, the distribution of phenotype sizes (S_3) is again very close to a log-normal law. Here $\mu = 5.604$ and $\sigma = 1.838$. The log-normal fit is worse than in **(a)** because there is a small *bump* on the right part of the distribution, where larger phenotypes are —due to the over presence of two-gene phenotypes (see text). **(d)** The rank distribution again shows a long tail. Only 300 phenotypes in $\mathcal{P}_3$ represent almost 99% of all genotypes. The remaining $26,000$ phenotypes are extremely rare in comparison.

However, the distribution of phenotypes is hardly even among genotypes (see Figure 3.2). As was the case in RNA, HP proteins, regula-

tory networks and so on, the distribution is highly skewed. In both cases, the distribution can be empirically fitted to a log-normal distribution, with probability density function

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{\sigma^2}\right),$$

where μ is the mean and σ is the standard deviation of the normally distributed logarithm of the variable. We obtained the parameters empirically from the log-transformed data (Figure 3.2**a** and **c**). The fit in the $g = 3$ case is worse due to a small *bump* in the right part of the distribution, where the larger phenotypes of $\mathcal{P}_3$ are. We will discuss this bump later on.

It is remarkable that both distributions somehow resemble a log-normal fit, because this is what has been found for RNA (Dingle et al., 2015) and for simple mathematical models completely unrelated to toyLIFE (Manrubia and Cuesta, 2017). These results point to a fundamental law underlying the distribution of phenotype sizes for general genotype-phenotype maps.

In both the $g = 2$ and $g = 3$ case, the rank distributions (Figure 3.2**b** and **d**) show a long tail, confirming that, indeed, while few phenotypes are very abundant, most of them are rare. In the $g = 3$ case this is especially striking, since 300 phenotypes in $\mathcal{P}_3$ represent nearly 99% of all genotypes —which means that the remaining 1% is represented in $\sim 26,000$ phenotypes!

All phenotypes in $\mathcal{P}_2$ are also found in $\mathcal{P}_3$: we can always add a gene that does not fold into any toyProtein to a viable two-gene genotype. A pertinent question, therefore, is how abundant these phenotypes are in three-gene genotype space. This is represented in Figure 3.3. In Figure 3.3**a**, we represent the size of a phenotype in $g = 2$ space (S_2) versus its corresponding size in $g = 3$ space (S_3), for each phenotype in $\mathcal{P}_2$. The Figure also shows a power-law fit, $\log_{10} S_3 = 6.064 + 0.986 \log_{10} S_2$, corresponding to $S_3 = 10^{6.064} S_2^{0.986} \approx 10^6 S_2$, a linear fit. This means that the size ordering between these phenotypes does not change when exploring genotypes with one more gene. The goodness of the fit is further shown in Figure 3.3**b**, which represents the histogram of values of $\log_{10}(S_3/S_2)$. The distribution is concentrated around its mean, 5.996, very close to the value 6.064 obtained in Figure 3.3**a**. This second result confirms that the size of $\mathcal{P}_2$ phenotypes in $g = 3$ space is equal to their corresponding size in $g = 2$

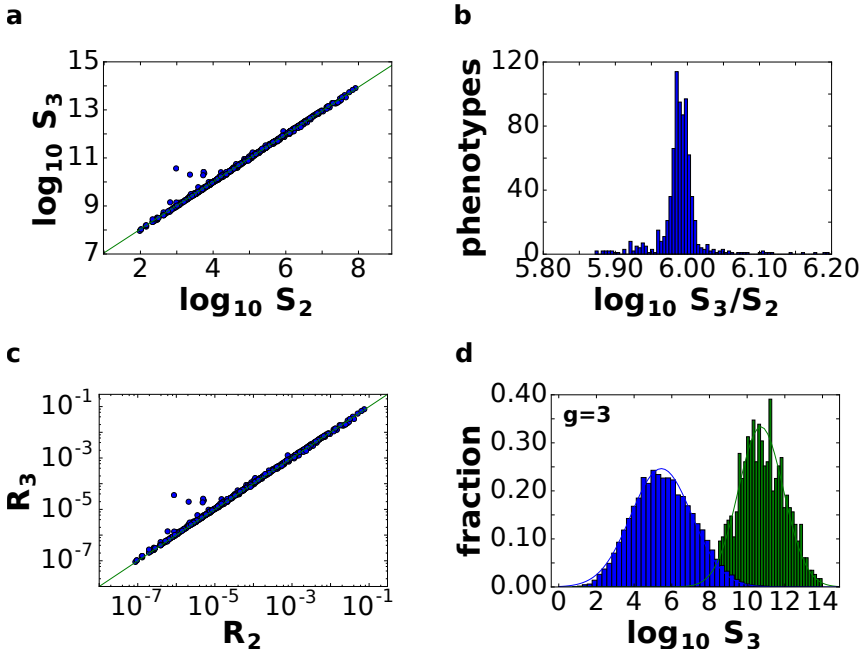


Figure 3.3: Two-gene phenotypes dominate phenotype space in the three-gene case. (a) The 775 phenotypes belonging to $\mathcal{P}_2$ also appear in $\mathcal{P}_3$. This figure represents the corresponding size of each phenotype in both genotype spaces: S_2 and S_3 are, respectively, phenotype size in $g = 2$ and $g = 3$ space. Green line represents the linear fit $\log_{10} S_3 = 6.064 + 0.986 \log_{10} S_2$, which is close to the linear fit $S_3 \sim 10^6 S_2$. **(b)** Histogram of $\log_{10}(S_3/S_2)$ for each of the 775 phenotypes in $\mathcal{P}_2$. The mean of the distribution is 5.996. **(c)** Relative size of the 775 phenotypes in $\mathcal{P}_2$ (R_2) versus their relative size in $g = 3$ space (R_3) —computed as phenotype size divided by number of viable genotypes. Green line is $R_3 = R_2$. The close fit means that the phenotypes from $\mathcal{P}_2$ dominate phenotype space in $g = 3$ space. **(d)** Size distribution of phenotypes in $\mathcal{P}_3$, taking the 775 phenotypes in $\mathcal{P}_2$ and rescaling them — we have obtained the two histograms as if they came from independent distributions for clarity. The green histogram represents the phenotypes in $\mathcal{P}_2$, and the blue histogram the remaining $25,717$ phenotypes in $\mathcal{P}_3$. New log-normal fits are drawn: $\mu_3 = 5.449$, $\sigma_3 = 1.619$ (blue line), $\mu_2 = 10.730$, $\sigma_2 = 1.196$ (green line). Note that the log-normal fit for three-gene phenotypes is much better once we take into account the 775 phenotypes in $\mathcal{P}_2$. All fits in this and subsequent Figures have been done using the least squares method.

space times 10^6. Where does this factor come from? Recall that there are $2^{20} \sim 10^6$ toyGenes in toyLIFE. A factor of almost 10^6 between S_3 and S_2 means that we can add almost any toyGene to a given two-gene genotype, and the resulting phenotype will be the same: it will not interfere with the original function. This is a remarkable fact.

Moreover, if we look at the distribution of relative sizes of $\mathcal{P}_2$ phenotypes —computed as phenotype size divided by the total number of viable genotypes— in $g = 2$ and $g = 3$ (Figure 3.3**c**), we obtain a linear relationship again: $R_3 = R_2$. Which means that the relative size of the phenotypes in $g = 2$ space is very similar to the relative size they represent in $g = 3$ space. But the sum of the relative sizes in $g = 2$ space is equal to 1 —there are only 775 phenotypes in $\mathcal{P}_2$. Accordingly, the sum of relative sizes in $g = 3$ is close to 1 —actually, it is 0.9964. This means that the 775 phenotypes in $\mathcal{P}_2$ dominate the space of phenotypes in $g = 3$ space. Only special combinations of three toyProteins and three promoters will yield different phenotypes in $g = 3$ space: the rest will be extensions of two-gene genotypes with a third toyGene that does not interfere in their function. The vast majority of phenotypes in $\mathcal{P}_3$ —99.71% of them— is generated by rare combinations of toyProteins, that represent less than 0.5% of genotype space.

Finally, let us look again at the histogram of phenotype size distributions in $g = 3$ that we obtained in Figure 3.2**c**. We can re-compute the histogram taking the 775 phenotypes from $\mathcal{P}_2$ as a separate set from the remaining $25,717$ phenotypes in $\mathcal{P}_3 - \mathcal{P}_2$. If we compute the respective histograms for both sets, we obtain Figure 3.3**d**. In green we have represented the 775 phenotypes in $\mathcal{P}_2$. It is not surprising that their distribution follows a log-normal law again: it follows immediately from Figure 3.2**a** and from the linear relationship shown in Figure 3.3**a**. What is relevant, however, is that the *bump* we observed in Figure 3.2**c** is gone in the histogram of the remaining $25,717$ phenotypes (in blue). In a sense, it is as if both sets were somehow independent: one is formed by two-gene genotypes with a third, non-interfering toyGene, and the other is formed by all combinations of three toyGenes that express something new, that was not present before.

A relevant question now is how important these 775 phenotypes in $\mathcal{P}_2$ are for larger genotypes. Exhaustive sampling of genotype spaces larger

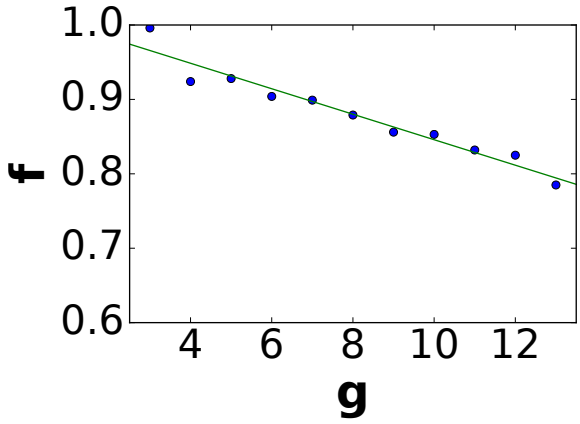


Figure 3.4: The dominance of two-gene phenotypes decays linearly with genotype size. For each g, we sample $10,000$ viable genotypes and compute their phenotypes, counting how many phenotypes belong to $\mathcal{P}_2$. We then represent the fraction f versus g. The data can be fitted to a linear function: $f = 1.02 - 0.02g$ (green line). The fraction of phenotypes belonging to $\mathcal{P}_2$ decays with g, albeit very slowly.

than $g = 3$ is out of our possibilities, but we can perform random samples of genotypes for different values of g and observe the fraction f of observed phenotypes that belong to $\mathcal{P}_2$. This is represented in Figure 3.4. Observe that, although this fraction decays linearly with gene size as $f = 1.02 - 0.02g$, the slope of the decay is very small, and therefore the fraction is always high —higher than 80% for $g \leq 13$. In other words, phenotypes in $\mathcal{P}_2$ continue to dominate phenotype space in toyLIFE for a moderate number of genotype sizes.

3.3 Neutral networks in toyLIFE

Point mutations in toyLIFE are easy to implement: they are changes in one of the nucleotides in one of the genes in the genotype. If the sequence has a H toyN in that position, then a mutation will change it to a P toyN, and vice versa. This definition of mutation induces a network structure in genotype space. Because of the enormous degeneracy of the genotype-phenotype map in toyLIFE, genotypes will, on average, have more than one neutral

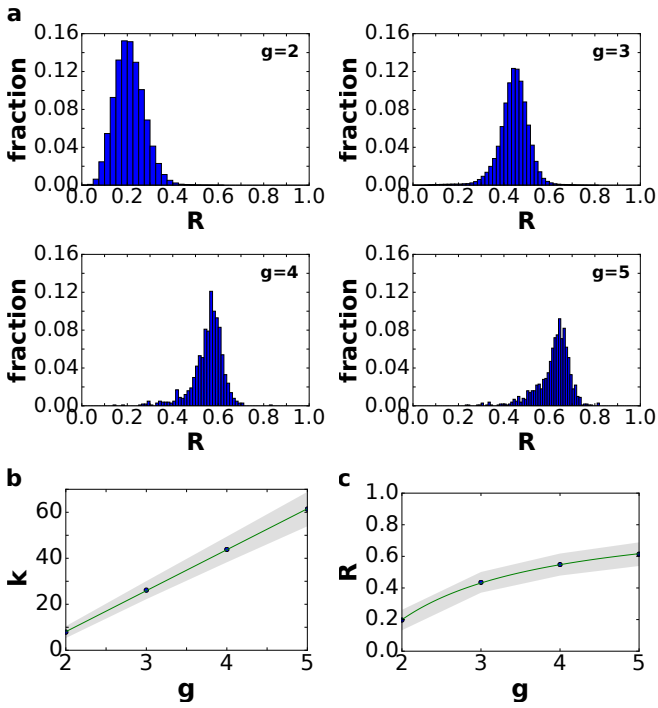


Figure 3.5: Genotypes in toyLIFE typically have a large number of neutral neighbors. **(a)** Distribution of robustness for genotypes for different values of g (gene number) for $g = 2$ to $g = 5$. Robustness is defined as the normalized degree of a node in the networks: $R = k/k_{\max}$, where k is the degree of a node in the neutral network, and $k_{\max} = 20g$ is the maximum degree in the network. Normalization allows for comparison of values among different genotype sizes. For $g = 2$ and $g = 3$, we sampled 10^7 genotypes, whereas for $g = 4$ and $g = 5$ we sampled $1,000$ genotypes. All distributions are unimodal, and more or less concentrated around the mean. Histograms for $g = 4$ and $g = 5$ are noisier because of reduced sampling size. **(b)** Average degree of nodes (blue circles) plus minus one standard deviation (gray area, empirically computed from the distributions in **(a)**) versus gene number g. The average degree $\langle k \rangle$ of a node grows linearly with gene number g, as $\langle k \rangle = -27.6 + 17.8g$ (green line). **(c)** Average robustness (blue circles) plus minus one standard deviation (gray area) versus gene number g. Robustness grows with gene number, and we can find an inverse relationship between both variables: $\langle R \rangle = 0.895 - 1.392/g$.

neighbor (Figure 3.5). Genotype-phenotype studies in the literature usually focus on robustness, defined as

$$R = \frac{k}{k_{\max}},$$

where k is the degree of a node in the neutral neighbor, and $k_{\max} = 20g$ is the maximum number of neighbors in the network. In other words, R is the normalized degree of a node. We can sample genotypes for different genotype sizes, represented by g (gene number), and plot the histogram of values of R (Figure 3.5**a**). tOYLIFE genotypes tend to be more robust as g increases. In fact, there is a linear relationship between g and $\langle k \rangle$, the average degree of a node in a neutral network (Figure 3.5**b**): $\langle k \rangle = -27.561 + 17.826g$. But $\langle R \rangle = \langle k \rangle / 20g$, so

$$\langle k \rangle \sim -27.561 + 17.8261g \iff \langle R \rangle 20g \sim -27.561 + 17.8261g \iff$$
$$\iff \langle R \rangle \sim -\frac{1.378}{g} + 0.891,$$

which is very close to the least-squares fit $\langle R \rangle = 0.895 - 1.392/g$, shown in Figure 3.5**c**. The linear relationship between $\langle k \rangle$ and g with slope 17.8 indicates that, on average, for every gene we add to a genotype, most mutations in the new gene will be neutral. This is consistent with the results obtained in Section 3.2, that showed that newly added genes tended not to interfere with the existing phenotype. These new genes will tend to work as *junk* in the sense that they will not affect the function and that mutations in their sequence tend to be neutral. We will see later on that *junk* genes also allow for more evolvability in tOYLIFE genotypes, suggesting interesting consequences for evolution.

Also, taking into account that $\mathcal{P}_2$ phenotypes dominate in $\mathcal{P}_g$ for $g \leq 13$ (Figure 3.4), we could estimate $S_g \sim 20^g S_2$, so $\log S_g \sim C + g \log 20$, where C is a constant. Combining this result with the linear relationship between g and $\langle k \rangle$, we obtain for tOYLIFE the linear relationship between $\langle k \rangle$ and $\log S$, that has been observed previously for RNA (Aguirre et al., 2011) (but see Figure 3.10 for a direct check of this relationship).

The average number of neutral neighbors (for any g) is larger than 1, so the neutral networks associated to these phenotypes will tend to have large connected components. As we mentioned in Chapter 1, for most

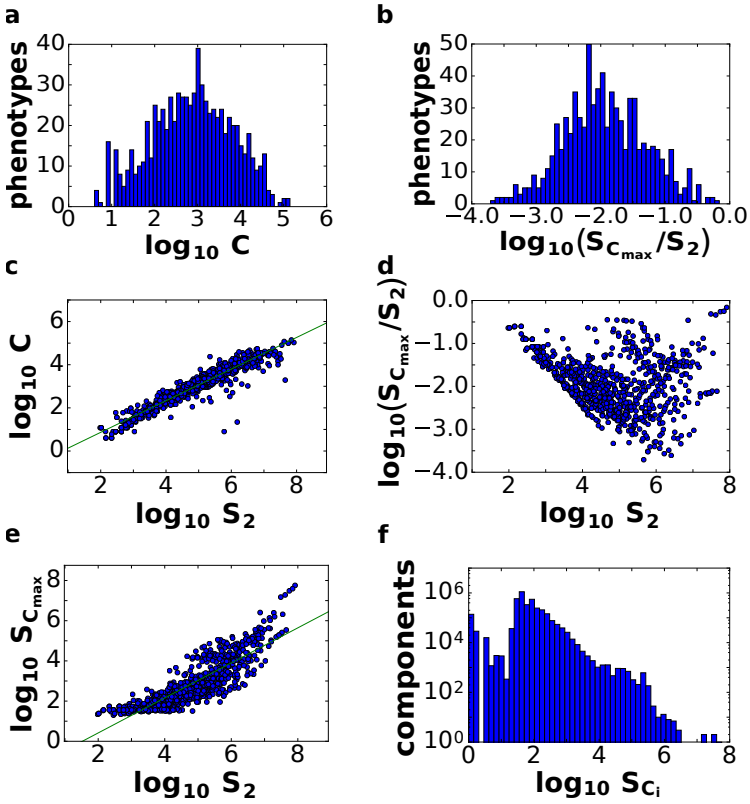


Figure 3.6: Neutral networks in toyLIFE are highly fragmented for $g = 2$. (a) For all 775 phenotypes in $\mathcal{P}_2$, we computed the number of connected components (C) of the associated neutral network. This figure represents the distribution of the decimal logarithm of C per neutral network. No single phenotype has less than 4 connected components. **(b)** For each neutral network, we take the maximal component Cmax and plot the distribution of the logarithm of its relative size — that is, the logarithm of $S_{C\text{max}}$ divided by S_2. **(c)** The size of the phenotype and the number of components are related via a power law: $C = 0.25 S_2^{0.7}$. **(d)** The relationship between the relative size of Cmax and the size of the phenotype is very noisy, but **(e)** there is a positive correlation between the absolute size of Cmax and the size of the phenotype. The green line represents the power law fit $S_{C\text{max}} = 0.05_2^{0.9}$. **(f)** Distribution of the logarithm of size of all connected components C_i in $g = 2$ space.

models of the genotype-phenotype map, neutral networks tend to have one giant component, although this is not always the case: RNA molecules of length 12 form neutral networks that are highly disconnected (Aguirre et al., 2011). Although network analysis is almost impossible for $g > 3$, as networks are enormous, for $g = 2$ we can perform network analyses on all 775 phenotypes exhaustively, and compute their connected components (Figure 3.6). We observe that most phenotypes are distributed in highly fragmented neutral networks: the genotypes corresponding to a given phenotype cluster in many disjoint connected components (Figure 3.6**a**): the number of connected components C is never smaller than 4 and is usually much larger. Moreover, these connected components tend to be small: if we consider Cmax, the maximal component associated to each neutral network, its average relative size $S_{C\text{max}}/S_2$ is 0.033 (Figure 3.6**b**). Only 63 phenotypes have connected components that are larger than 10% the phenotype size —among these are the largest connected components in $g = 2$ space, including one giant network that contains $56,889,472$ nodes!

Large phenotypes tend to have a larger number of connected components, and we can find a relatively good power-law fit between the size of the phenotype S_2 and the number of components C: $C = 0.25S_2^{0.7}$ (Figure 3.6**c**). The relationship between S_2 and the relative size of Cmax is noisy (Figure 3.6**d**): smaller phenotypes have less connected components and therefore the relative size of the maximal component is high. As the number of components increases, most of them tend to have equal, small sizes. However, the largest phenotypes with the greatest number of connected components also have the largest connected components, as we pointed out before, so there is a positive correlation between S_2 and the absolute size of its maximal component, $S_{C\text{max}}$. This last fact is represented in Figure 3.6**e**.

In short, there is a huge variation in the size of connected components in $g = 2$. We can plot the distribution of sizes of all connected components C_i —irrespective of the phenotype they belong to (Figure 3.6**f**). The average component size, S_{C_i}, is 301.4, but we can see from the histogram that the distribution has a long tail. Therefore, although most connected components are smaller than $1,000$ nodes — roughly 98.5%!— some of them are larger, reaching up to $\sim 10^7$ nodes. These results imply that navigability in $g = 2$ space is somewhat limited, and that the numerous properties

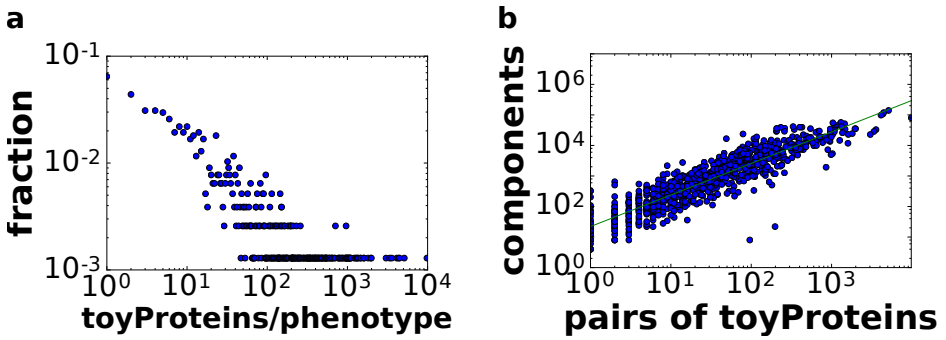


Figure 3.7: Most phenotypes in $\mathcal{P}_2$ are obtained by a small number of pairs of toyProteins. (a) Distribution of the number of pairs of toyProteins that generate a given phenotype. For example, if both $\{1,1\}$, $\{1,2\}$ and $\{3,4\}$ generate a given phenotype, there are 3 pairs of toyProteins that generate it. **(b)** Due to the HP model that underlies toyProtein folding, the more pairs of toyProteins are able to generate a given phenotype, the larger the phenotype and, because of the power-law relationship obtained in Figure 3.6**c**, the more connected components that will belong to the phenotype. The green line represents the power-law fit $C = 22.093P^{1.032}$.

granted by neutral networks, as commented in Chapter 1, will not be used to the greatest advantage.

The high disconnection in connected components is due to the HP model that underlies toyProtein folding. Any given phenotype in $\mathcal{P}_2$ will be obtained by some set of pairs of toyProteins. Figure 3.7**a** shows that this distribution is highly skewed, with a long tail: 28.64% of phenotypes in $\mathcal{P}_2$ are obtained by less than 10 pairs of toyProteins, while one phenotype is obtained by $9,808$ pairs of toyProteins. The problem, therefore, is not due to a small set of toyProteins associated to each phenotype. Rather, the cause of the disconnection between connected components is due to the HP model, because there are no neutral mutations among toyProteins (see Chapter 2). In other words, every mutation in a toyGene will yield a different toyProtein or will not fold, but will never generate the same toyProtein. Moreover, the connections between different toyProteins are scarce, thus difficulting the change from one to the other. As a result, the more toyProteins generate a given phenotype, the more connected components will belong to this phenotype (Figure 3.7**b**).

For $g \geq 2$, we can estimate the distribution of neutral networks in genotype space using neutral random walks: starting at a randomly chosen genotype, we perform a mutation on it. If the resulting mutant genotype belongs to the same neutral network —that is, it expresses the same phenotype— the mutation is accepted: the random walk continues when we mutate the new genotype again. If the mutant genotype does not belong to the neutral network, the mutation is rejected, and we try to find a new neutral neighbor for the original genotype. This process will not work if the starting genotype does not have neutral neighbors, but looking at Figure 3.5 it is easy to see that this event will happen very rarely. Once

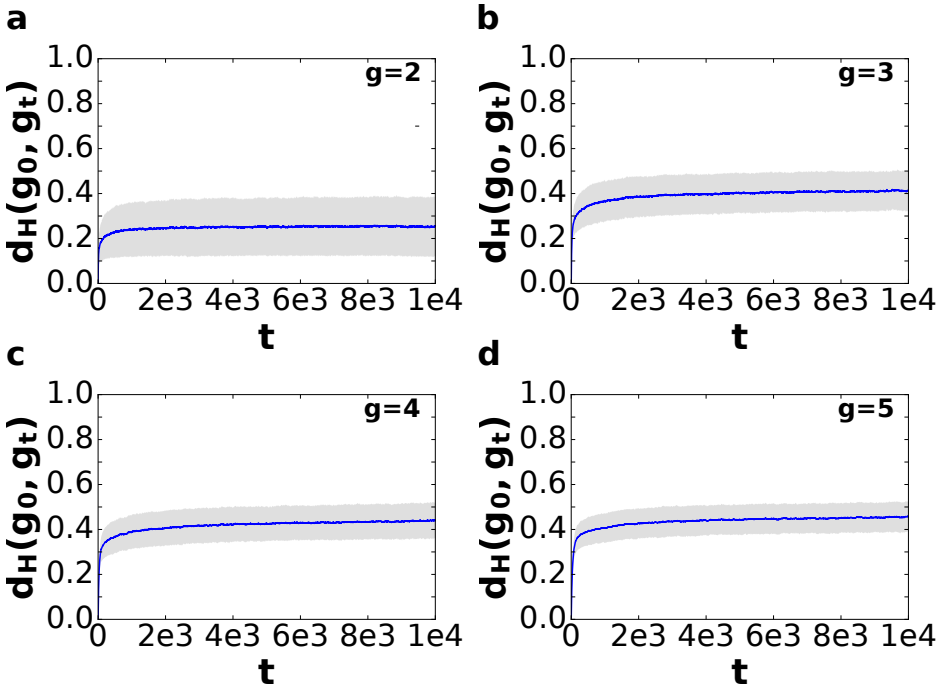


Figure 3.8: Neutral networks in toyLIFE span a large fraction of genotype space (1). For each genotype size, from $g = 2$ to $g = 5$, we performed $1,000$ neutral random walks starting at randomly chosen genotypes. The length of the random walks was $10,000$ time steps. The figure represents the average Hamming distance $\langle d_H \rangle$ (blue line) between the genotype visited at time t, g_t, and the original genotype g_0, plus minus one standard deviation (grey area), empirically obtained from the data.

we start the random walk, we can keep count of the separation between the genotype at time step t and the original genotype. For that end, we need a metric: we will use the relative Hamming distance, briefly mentioned in Chapter 1. For two genotypes g_1 and g_2, the Hamming distance is

$$d_H(g_1, g_2) = \frac{1}{20g} \sum_{i=1}^{20g} \delta(n_{1,i}, n_{2,i})$$

where $n_{j,i}$ is the toyN of the genotype g_j at position i and $\delta(n,m)$ is Kronecker's delta, which is 1 if $n = m$ and is 0 otherwise. The normalization factor $20g$ ensures $0 \leq d_H \leq 1$, allowing for comparison between different g. For example, if g_1 =HPPH and g_2 =HPHP, $d_H(g_1, g_2) = 0.5$ because the last two toyN differ.

We performed $1,000$ neutral random walks of length $10,000$ for genotype sizes $g = 2$ to $g = 5$ (Figure 3.8). That is, for each g, we randomly sampled $1,000$ genotypes, and performed the random walk process described above. At each time step t, we computed $d_H(g_0, g_t)$, the Hamming distance between the original genotype g_0 and the genotype visited at time t, g_t. $d_H(g_0, g_t)$ is a random variable for each t, and so we can compute its average and standard deviation, and plot them (Figure 3.8). If there were no restrictions to the nodes that can be visited in a random walk, we would expect $d_H(g_0, g_t) \to 0.5$ when $t \to \infty$. In other words, if there are no restrictions, the correlation between g_0 and g_t is lost when t grows, and the distance between them tends to the value it would have, on average, if we randomly picked two genotypes from the network. Thus, the evolution of $d_H(g_0, g_t)$ is a good measure of the size and distribution of neutral networks in genotype space.

For $g = 2$, $\langle d_H(g_0, g_t) \rangle \to \sim 0.25$ when $t \to \infty$, implying that networks are not very large. Considering that the total genotype space has diameter 40, this means that the average distance between the initial genotype and the final one is close to 10. This is not a very high value, and it is consistent with our previous analysis showing that neutral networks in $g = 2$ tend to be fragmented and small.

For $g > 2$, $\langle d_H(g_0, g_t) \rangle \to \sim 0.4$ when $t \to \infty$, which implies that the fragmented networks of $g = 2$ space are becoming more connected as g grows, facilitating the navigability in genotype space. This suggests that

neutral networks for $g > 2$ span large fractions of genotype space, a result consistent with other unrelated models of the genotype-phenotype map.

A different way to estimate the diameter of a neutral network is to perform neutral random walks in which we force $d_H(g_t, g_{t+1}) > d_H(g_{t-1}, g_t)$. That is, additionally to imposing that the mutation is neutral in order to accept it, we also need it to increase the distance to the original genotype. The process is computed as follows: we randomly choose a genotype, and perform mutations on it, increasing the distance every time step, until this distance can increase no longer —if after a large number of trials we cannot find a neutral mutant that is farther apart from the original genotype, we stop the process. We will term the final distance obtained in such random walks d_∞. For $g = 2$ and $g = 3$ we randomly sampled $10,000$ genotypes, whereas for $g = 4$ and $g = 5$ we sampled $1,000$ genotypes (Figure 3.9**a**). Consistent with previous results, random walks did not get very far in $g = 2$ space and the average final distance $\langle d_\infty \rangle$ is ~ 0.2.

For $g > 2$, the final distance d_∞ increases. This result confirms the previous discussion that navigability in these genotype spaces is enhanced. For $g = 3$, $\langle d_\infty \rangle$ is a little over 0.5, while for $g = 4$ and $g = 5$ it gets to 0.6 and 0.7, respectively. In fact, the growth of $\langle d_\infty \rangle$ with g is very similar to the growth of $\langle R \rangle$ obtained in Figure 3.5**c**. In that case, we had $\langle R \rangle = 0.895 - 1.392/g$. Here it is $\langle d_\infty \rangle = 0.965 - 1.354/g$ (Figure3.9**b**). Unsurprisingly, the similarity of the fits implies a linear relationship between $\langle d_\infty \rangle$ and $\langle R \rangle$: $\langle d_\infty \rangle = 0.094 + 0.972\langle R \rangle$ (Figure 3.9**c**), very close to the identity. This result has several implications. First, as g grows, neutral networks are more and more connected, and they span larger fractions of genotype space. It is easier to get from one extreme of the genotype network to the other without changing the phenotype. Secondly, this increased connectivity is due to the increase in robustness: the robustness of a genotype is a good predictor for the size of the connected component it belongs to. This can be easily explained in light of our previous discussion on robustness. Adding a new gene to a genotype will give it 18 or so new neutral mutations with which to play (Figure 3.5**b**). Because this new gene will not interfere with the phenotype with a high probability, it follows that we can mutate most of its nucleotides, one by one, getting farther away from the original genotype. In other words, new genes in tQYLIFE allow for increased navigability of genotype space because they are mostly *junk*

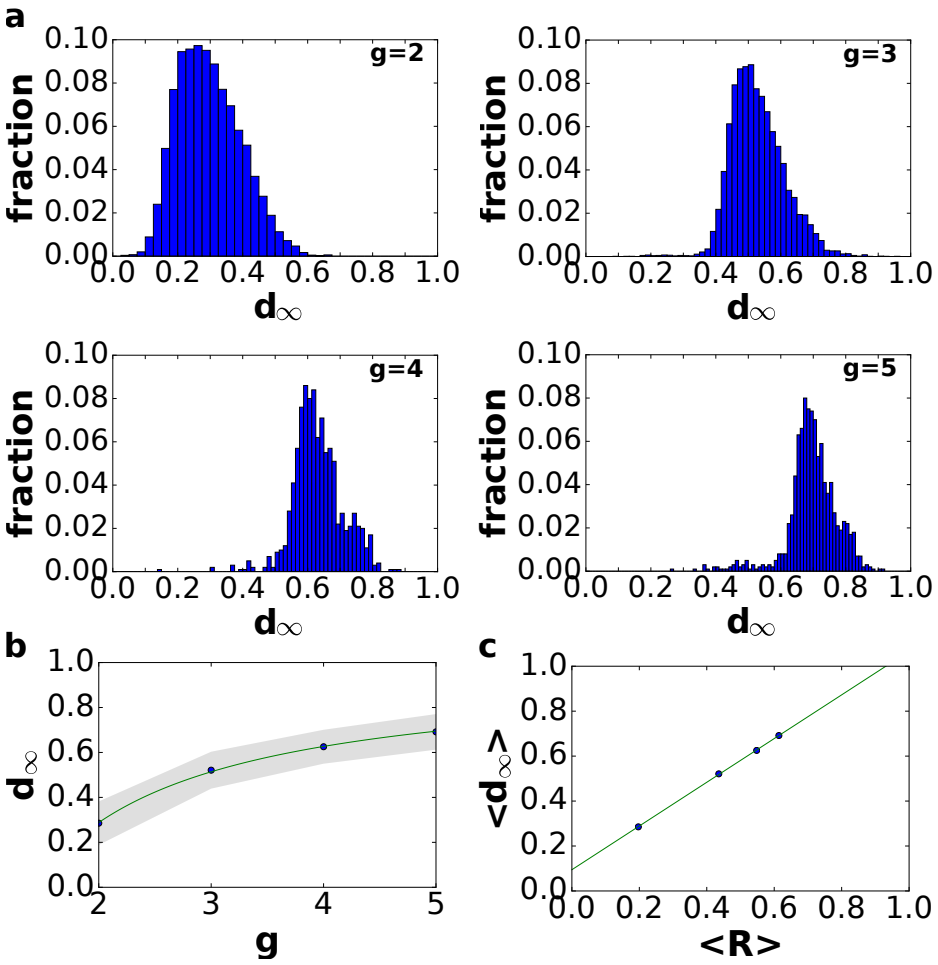


Figure 3.9: Neutral networks in toyLIFE span a large fraction of genotype space (2). (a) We performed $10,000$ (for $g \leq 3$) or $1,000$ (for $g > 3$) neutral random walks, forcing them to increase the Hamming distance to the original genotype. We stopped when the random walk could get no farther. **(b)** There is an inverse relationship between g and $\langle d_\infty \rangle$: $\langle d_\infty \rangle = 0.965 - 1.354/g$ (green line). The blue circles represent $\langle d_\infty \rangle$, whereas the gray are is plus minus one standard deviation. **(c)** There is a linear relationship between $\langle d_\infty \rangle$ and the average robustness of the genotypes as obtained in Figure 3.5**c**, given by: $\langle d_\infty \rangle = 0.094 + 0.972 \langle R \rangle$ (green line), very close to the $\langle d_\infty \rangle = \langle R \rangle$ fit.

genes. As we will see later on, this property will have important consequences for evolvability.

The fact that robustness is a good predictor for the size of a genotype's connected component can be combined with the positive correlation between the logarithmic size of a genotype and the size of its largest connected component (Figure 3.6**d**) to deduce the relationship between the logarithm of phenotype size and phenotypic robustness, as had been observed before in other models (Aguirre et al., 2011).

Phenotypic robustness is defined as the average of genotypic robustness for all genotypes belonging to a phenotype $\mathcal{P}_i$, that is

$$R_{\mathcal{P}_i} = \frac{1}{|\mathcal{P}_i|} \sum_{g \in \mathcal{P}_i} R_g,$$

where $|\mathcal{P}_i|$ is the number of genotypes belonging to $\mathcal{P}_i$. For $g = 2$ and $g = 3$ we sampled 10^7 genotypes and computed their robustness. We then

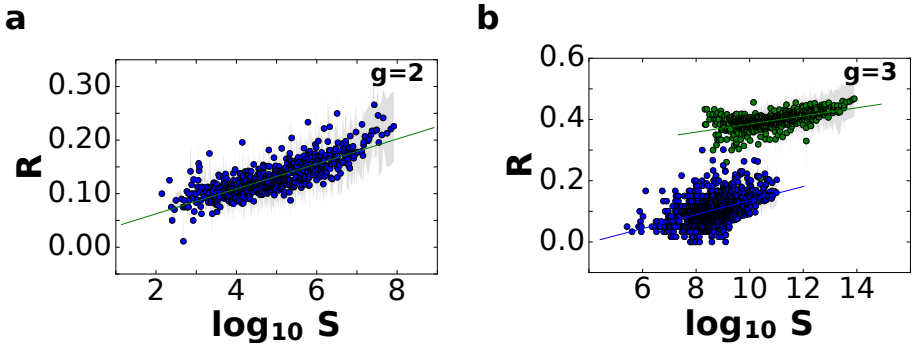


Figure 3.10: Phenotypic robustness is linearly related to the logarithm of phenotype size. For $g = 2$ and $g = 3$, we sampled 10^7 genotypes and computed their robustness. Then we assigned each of them to their corresponding phenotypes, and estimated phenotypic robustness, the average robustness for all genotypes belonging to a given phenotype (see text). **(a)** Phenotypic robustness in $g = 2$ versus the logarithm of phenotype size. The green line represents the power-law relationship $R_p = 1.037\, S_2^{0.023}$. **(b)** For $g = 3$, we separated those phenotypes belonging to $\mathcal{P}_2$ (green circles) from the rest (blue circles). Both sets show a power law relationship between phenotypic robustness and phenotypic size: $R_p = 1.790\, S_3^{0.013}$ for phenotypes in $\mathcal{P}_2$ (green line), and $R_p = 0.805\, S_3^{0.023}$ for the remaining $25,717$ phenotypes (blue line).

assigned each genotype to its corresponding phenotype and averaged the values for all genotypes belonging to that phenotype. Note that this procedure samples large phenotypes more often. For $g = 2$, we find a good fit to a linear relationship between the logarithm of phenotype size and estimated phenotypic robustness (Figure 3.10**a**). For $g = 3$, we identified those phenotypes belonging to $\mathcal{P}_2$ (being the largest, they were sampled the most) in green, and the rest in blue (Figure 3.10**b**). Separate linear relationships between logarithm of phenotype size and phenotypic robustness are drawn. Amazingly, the two sets of phenotypes cluster in two different groups, confirming once more the idea that these two sets are qualitatively different. Phenotypes belonging to $\mathcal{P}_2$ are much more robust —indeed, most of the histogram in Figure 3.5**a** is due to them—, as a result of them having one spare *junk* gene. Nevertheless, the linear relationship between the logarithm of phenotype size and phenotypic robustness is kept in both sets, hinting at a general property of genotype-phenotype maps.

3.4 Robustness and position in genotype

Instead of considering the degree of a node in the neutral network, we can focus on the neutrality of a given position. For a given sequence, the position $i = 1, \ldots, 20g$ can either be neutral or not —that is, when we mutate that position, we can get a new genotype with the same phenotype or not. We can thus define the random variable

$$r_i = \begin{cases} 1 & \text{if } i \text{ is a neutral position,} \\ 0 & \text{otherwise.} \end{cases}$$

Being a random variable, we can sample it and estimate its average: if there are differences, we will get insights into the details of toyLIFE as a model of the genotype-phenotype map. This is what we have done in Figure 3.11, for genotype sizes $g = 2$ to $g = 5$. We sampled 10^6 genotypes for $g = 2$ and $g = 3$, and 10^4 genotypes for $g = 4$ and $g = 5$, and computed r_i for every $i = 1, \ldots, 20g$ and every genotype. However, the order of the toyGenes does not matter in toyLIFE by construction —implying $\langle r_i \rangle = \langle r_{i+20h} \rangle$, for any $h \in \mathbb{N}$—, so we are interested in the values of robustness inside each toyGene. This is why in Figure 3.11 we only show the average values $\langle r_i \rangle$ for $0 \leq i < 20$. Note that the promoter regions tend to be more robust than

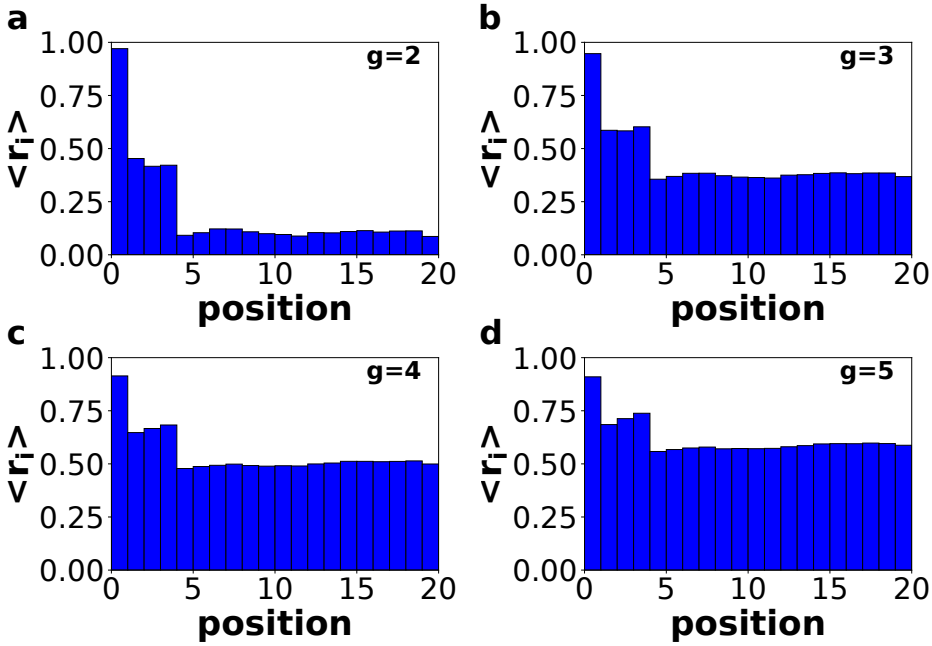


Figure 3.11: Different positions in the genome have different neutralities. We sampled 10^6 genotypes for $g = 2$ and $g = 3$ and 10^4 genotypes for $g = 4$ and $g = 5$ and measured r_i for $i = 1, \ldots, 20$ —every relevant position in the genome, since the order of the toyGenes does not matter. For each i we then computed $\langle r_i \rangle$ and plotted them versus the position. Note the high robustness of the first position in the promoter region, and the low robustness in the coding regions.

the coding regions. This is due, partly, to the lack of robustness in the HP model underlying the toyProtein folding (as discussed in Chapter 2). However, note that the superposition of regulatory and metabolic levels of the phenotype makes the average robustness of the coding regions grow, in spite of the HP model. For $g = 4$ and $g = 5$, the average robustness in these regions goes as high as 0.5. Remember that these genotypes tend to have non-interfering, junk toyGenes, that increase overall robustness as g grows.

Inside the promoter regions, the first position is particularly robust. This means that the regulatory changes it induces are mostly unseen. This may be due to two reasons: either changes in the first position of the promoter region do not affect the regulatory function —the logic function de-

termined by the interactions among toyProteins— or changes in the regulatory function rarely alter the metabolic phenotype. A simple test of these hypotheses is the following: for each position in the promoter region, we sampled $10,000$ genotypes of size $g = 3$. We then mutated that position and computed the new regulatory cycle and the new phenotype. From all $10,000$ mutations in the first position, 40.11% were neutral in both the regulatory and the metabolic sense. 54.25% affected the regulatory cycle but did not affect the metabolic phenotype and the remaining 5.64% changed both —this means that the measured robustness for the first position in this sample was 94.36%. For the rest of the positions, 26.81% of the mutations did not alter either the regulatory cycle or the metabolic phenotype, 32.37% changed the cycle but not the phenotype, and 40.82% changed both —a robustness of 59.18%, coherent with what we observed in Figure 3.11. In other words, for the first position only 9% of the mutations that affected regulation had any effect on the phenotype. Besides, there were many mutations —40% of them— that did not affect the regulatory function at all. For the rest of the positions, however, the number of mutations that altered the regulatory function was higher —73% of them— and, among these, roughly 55% had an effect on phenotype as well. So both reasons posited above apply: not only the number of mutations affecting regulatory function is lower in the first position of the promoter region, but when these mutations do alter the regulatory function, they rarely change the phenotype.

The lower robustness of coding regions, compared to promoter regions, is correlated with a higher evolvability, as will be discussed in the next Section.

Besides measuring the $\langle r_i \rangle$, we can compute the correlations between the r_i to study further relationships (Figure 3.12). The correlation matrix $\mathbf{C}$ has components

$$c_{ij} = \frac{\mathrm{Cov}(r_i, r_j)}{\sigma_i \sigma_j} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\sigma_i \sigma_j},$$

where $\sigma_i = \sqrt{\mathrm{Var}(r_i)}$. In this Figure we can observe, for $g > 2$, a high correlation inside the coding region and between the promoter region and the coding region, meaning that when a toyGene is robust, it tends to be robust everywhere, and vice versa. There are no special positions inside

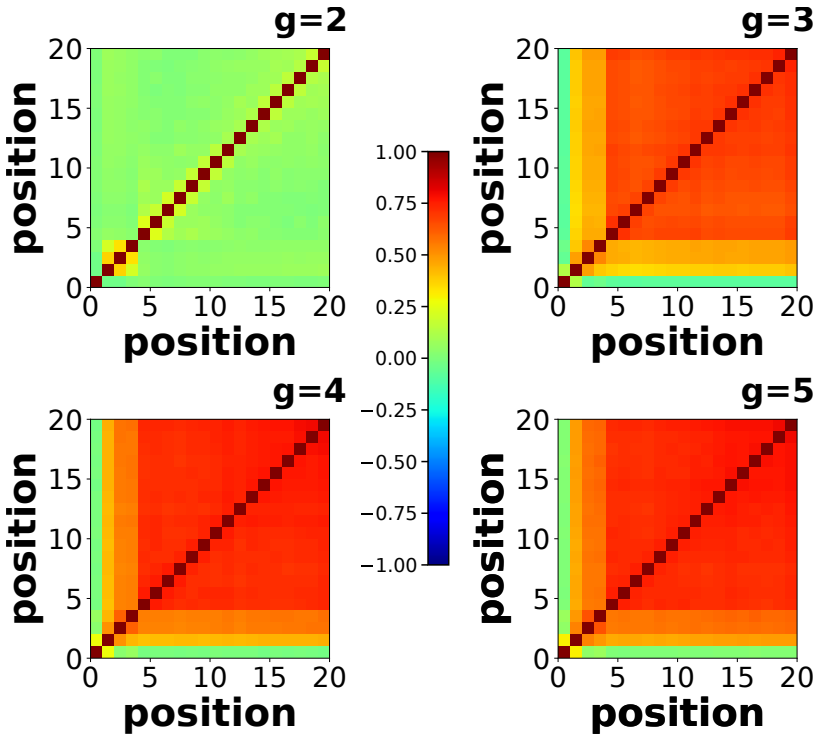


Figure 3.12: Correlation matrices of robustness per position for different genotype sizes. Using the same data as in Figure 3.11, we computed the correlation between r_i and r_j for all $1 \leq i, j \leq 20$ (see text). Note that the correlation is much higher inside the coding region.

the coding region (as can be clearly seen in Figure 3.11). Finally, note that the first position in the promoter region is uncorrelated with the rest of the positions, as it is always highly robust.

3.5 Accessibility and Evolvability

So far, we have limited our discussion of the properties of the genotype-phenotype map in toyLIFE to the size and distribution of neutral networks, without paying any attention to the connections between them. In this

section we will focus on this question, which is no other than the study of evolvability, or how accessible phenotypes are.

As we mentioned in Chapter 1, neutral networks in most models of the genotype-phenotype map tend to be highly interwoven, such that connections between them are very common. The Vienna group, led by Peter Schuster, described a property of RNA neutral networks, called *shape space covering* (Schuster et al., 1994; Grüner et al., 1996b). This means that we will be able to find most common phenotypes a few mutations away from any given genotype. We checked for the existence of this property in toyLIFE. In order to do so, we sampled 100 genotypes for $g = 2$ and $g = 3$ and computed the phenotypes of all neighbors at distances 1 to 8. We observed how many of the 300 most common phenotypes appeared in this set of neighbors. The results are shown in Figure 3.13. For both $g = 2$ and $g = 3$, for most sampled genotypes the number of phenotypes discovered after 8 mutations was close to 300. For $g = 2$, the average number of phenotypes was barely over 278, while for $g = 3$ this number was close to 293. This implies that toyLIFE also shares the *shape space covering* property:

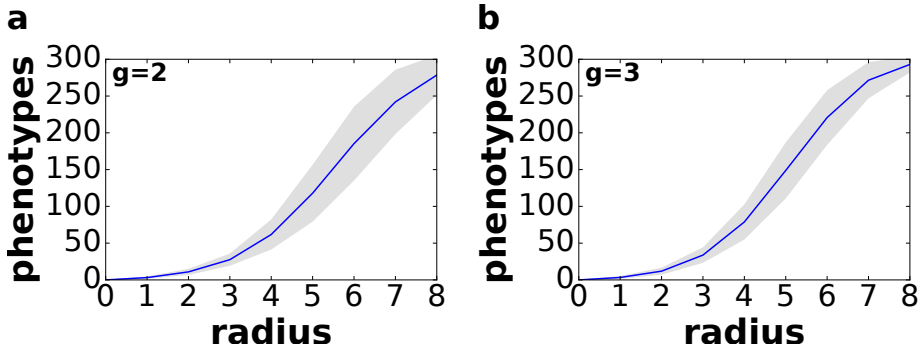


Figure 3.13: Shape-space covering in toyLIFE. We say that a genotype-phenotype map has the shape space covering property if, given a phenotype, we only need to explore a small radius around a sequence belonging to that phenotype in order to find the most common phenotypes. We tested this property in toyLIFE by sampling 100 genotypes for $g = 2$ and $g = 3$, and computing the phenotypes for all genotypes in a radius of distance 8 around that given genotype. The results are consistent with shape space covering. The figure shows the average (blue line) plus minus one standard deviation (gray area).

most phenotypes are just a few mutations away from any given phenotype. Observe, however, that for $g = 2$ this means a higher relative distance — remember that the diameter of this network is 40— and that the number of phenotypes discovered at that distance is lower in comparison.

Shape space covering means that phenotypes are easily accessible from each other through a few number of mutations. A relevant detail in the metabolic genotype-phenotype map in $\mathtt{toy}$LIFE is that this accessibility is due only to mutations in toyProteins. If we mutate only the promoters, the number of visited phenotypes is never larger than 2, independently of the distance. In $g = 2$, we can give a clear explanation to this peculiarity: of the $135,318$ pairs of toyProteins that yield a metabolic function, only 16 are associated with two phenotypes —they can yield two different metabolic phenotypes when combined with different promoters. The rest belong to only one phenotype (remember, however, that a given phenotype can be related to many pairs of toyProteins, Figure 3.7). Changing only the promoters will not affect the metabolic function, and will not help in finding new phenotypes. This is consistent with the high robustness of promoter sites: non-neutral mutations in the promoter regions are actually lethal mutations.

Interestingly, the metabolic phenotype defined for $\mathtt{toy}$LIFE shows shape space covering, while the underlying model behind this function, the HP protein folding model, does not show it (Bornberg-Bauer, 1997; Ferrada and Wagner, 2012). As was the case with robustness, adding new levels of expression to a phenotype makes it more evolvable. This points, again, to a fundamental property of biological systems, in which superposed levels of complexity allow for a more robust, more evolvable phenotype.

Another way to study evolvability is to measure, directly, the connections between different phenotypes. We say two phenotypes are connected if there is one mutation connecting two genotypes belonging to each phenotype. The network of phenotypes thus created is undirected and weighted —the weight of an edge between two phenotypes is the sum of all edges connecting two genotypes belonging to each phenotype. This network admits auto-loops, and the weight of these edges is twice the number of edges connecting genotypes belonging to a given phenotype, also equal to the sum of the degrees of all the nodes belonging to a phenotype, in turn equal to the phenotype's relative robustness times the maximum

degree. For $g = 2$, where we can compute the whole network of genotypes with their corresponding phenotypes, we can build this phenotype network exhaustively. The network is not entirely connected: there is a giant component that includes 767 nodes out of the 775, and six additional tiny components, five of them with just one node and the remaining one with three nodes. So, for $g = 2$, some phenotypes will be unreachable by

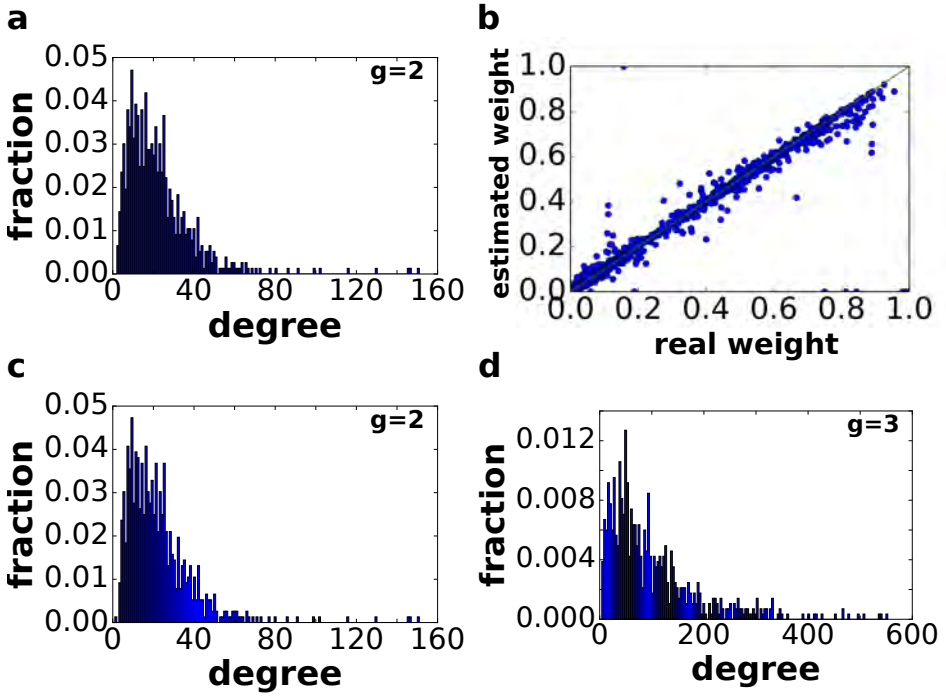


Figure 3.14: Connections between phenotypes in toyLIFE. (a) Degree distribution of the phenotype network in $g = 2$. Two phenotypes are connected if there is at least one genotype belonging to the first that can mutate into another genotype belonging to the second phenotype. The average degree is 22.134. **(b)** Estimated relative weight between phenotypes versus actual relative weight. Estimation performed by a random walk among all viable genotypes in $g = 2$. Length of the random walk is 10^9. The correlation between both variables is 0.978. **(c)** Estimated degree distribution from the previous random walk, for $g = 2$. **(d)** Estimated degree distribution for $g = 3$, using a random walk among genotypes belonging to phenotypes in $\mathcal{P}_2$.

point mutations, unless evolution starts in them. Additionally, the results show that the average degree is low, just 22.1, with a standard deviation of 17.3 (Figure 3.14**a**). The maximum degree is 151 and the minimum is 2. The largest weights are always those of the auto-loops —that is, the majority of connections in the genotype network do not change phenotype, consistently with our previous discussion on robustness. In fact, because not all phenotypes are equally large, we can compute the weighted average degree of the network —giving more weight to larger phenotypes. The result is an average degree of 54.0, illustrating that larger phenotypes are more connected than the average.

For $g = 3$, we can't build the phenotype network exhaustively. We will resort to a numerical approximation, in order to estimate the degrees of the nodes and their relative weights. Suppose we perform a random walk over all viable genotypes —jumping among them without any additional rule. If all genotypes are connected to each other —given our results for $g = 2$, this does not seem a terrible assumption— then we expect that, as the length of the random walk tends to infinity, every phenotype is visited proportionally to its size, and that the visits from one phenotype to another are proportional to the actual number of connections between them. The average number of visits (per time step) from phenotype i to j as time tends to infinity will be the same as the number of connections between phenotypes i and j, divided by the total number of connections leaving i.

We can check if this approach is accurate by performing the random walk on $g = 2$ space, for which we have the actual connection data. We performed a random walk starting at a randomly chosen genotype for 10^9 time steps. The relative weights computed by this method are close to the actual weights, as shown in Figure 3.14**b**. The correlation between both variables is 0.978: the outliers correspond to small phenotypes, which are hardly visited in the random walk. Figure 3.14**c** shows the estimated degree distribution, and it is obvious that is very similar to the one obtained from the actual network in Figure 3.14**a**.

Having made sure that this approach works, we repeated it on $g = 3$ space, again with a random walk of length 10^9 time steps. We restricted the random walk to the 775 phenotypes in $\mathcal{P}_2$: we wanted to study how the addition of one gene altered the connections between these phenotypes. When one mutation left this set of phenotypes, we considered it as lethal.

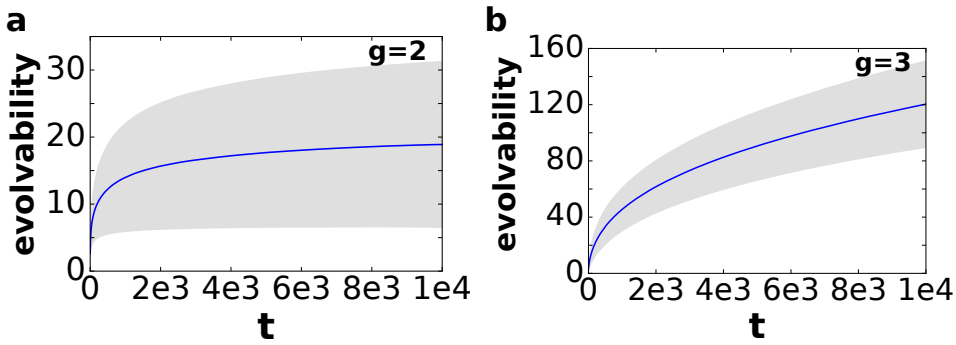


Figure 3.15: Evolvability in toyLIFE. For $g = 2$ and $g = 3$, we measured the cumulative number of phenotypes in the neighborhood of a neutral random walk —evolvability— for $10,000$ genotypes. The figure shows the average evolvability (blue line) plus minus one standard deviation (gray area), for random walks of length $10,000$. Note that evolvability is much higher for $g = 3$.

The results obtained show that all phenotypes in $\mathcal{P}_2$ now belong to one giant component —there is one phenotype that does not appear in the sample, but did belong to the giant component in $g = 2$ space, so it must belong to it in $g = 3$ space. The average degree is higher, 101.1, with a standard deviation of 90.3 (Figure 3.14**d**). The maximum degree is 553, and the minimum is 4. As we can see, the degree distribution is much wider, and the connectivity between phenotypes has been greatly enhanced. This is again due to the *junk* genes added to the genome. They do not only increase robustness, but also allow for increased connections between phenotypes. The weighted average degree is 333.3, again showing that larger phenotypes are much more connected than smaller ones.

This increased connectivity is also seen in Figure 3.15. Wagner (2011) estimates evolvability as the number of new phenotypes discovered in a neutral random walk along a neutral network. In Figure 3.15 we have performed those simulations for $10,000$ genotypes in $g = 2$ and $g = 3$ space. The results show that evolvability is much higher in $g = 3$ space: while the number of discovered phenotypes almost plateaus in $g = 2$, growing very slowly to the weighted average degree of 54.0, the number grows quickly in $g = 3$, and much higher than in $g = 2$ —again, this is due to the higher average degree in $g = 3$ space.

Increased navigability of genotype space allows for increased connectivity between phenotypes, thus enhancing evolvability. In other words, *junk* genes have creative potential, in the sense that they allow populations to explore a given neutral network, and then encounter new, unexplored phenotypes. This property is shared with most models of the genotype-phenotype map, but it is nice to see that tOYLIFE can show the evolutionary potential of new genes.

3.6 Summary

In this chapter, we have explored the properties of the metabolic genotype phenotype map in tOYLIFE. We have shown that, along with many other models of the genotype-phenotype map, tOYLIFE shows a high degeneracy, a very skewed distribution of phenotype sizes, the existence of neutral networks, growing robustness with genotype size, and high evolvability and accessibility, including shape space covering.

The fact that both RNA and tOYLIFE show this latter property, as well as the log-normal distribution of phenotypes (Figure 3.2, Dingle et al. (2015)), although the two models have almost nothing in common, points to a deeper reason underlying the generalities of the genotype-phenotype map. There must be some general principle built into these models that make them generate very similar outputs, in spite of the fact that their model objects do not necessarily share physic-chemical properties. The relationship between robustness, phenotype size and the log-normal distribution of phenotype sizes is one aspect of this general principle. If we can find a whole picture binding all these properties together, and if this principle can be extended to more general conditions, we may be able to make some interesting predictions for real systems, that escape our computational and modelling abilities at the moment.

On a different note, and more specific to tOYLIFE, we have seen how genotypes are prone to include *junk* genes as the size of the genotype grows. Most new genes added to a genome will not alter its function in a relevant way. However, this junk is not completely inert. First, it increases robustness, by increasing the number of neutral mutations that a genotype can admit. Secondly, evolvability is also increased. Sometimes out of the junk comes a new function, that has been mutating without re-

strictions. This is very interesting because it points out to relevant features shared by most eukaryote genomes, such as the abundance of introns, or the high expanses of seemingly non-functional DNA. If this non-functional DNA also enhances robustness and evolvability in living cells, then natural selection could act to preserve it.

Finally, it is remarkable that toyLIFE shows enhanced robustness and evolvability compared to the HP model on which it is based. It would seem that adding levels of complexity to the genotype-phenotype map makes evolution more robust and evolvable. This result ties in with Waddington's concept of canalization (Waddington, 1942) and, in general, with all the progress being made in the evo-devo field. However, we also know that complex genotype-phenotype maps come with their share of constraints and trade-offs. The final outcome of this contradicting tendencies is yet to be fully studied.

4

Functional promiscuity in models of the genotype-phenotype map

"Never put all your eggs in just one basket."

Traditional Spanish saying

Functional promiscuity, the property to carry out more than one function, has been well explored for enzymes. However, this property is not unique to them and can also be studied in other molecular models of the genotype-phenotype map. In this chapter, we extend the definition of functional promiscuity to RNA secondary structure, Boolean gene regulatory networks (GRNs) and toyLIFE. We will show that promiscuity is widespread in these models, and that it allows for the discovery of a large number of new phenotypes.

In the event of an environmental change, functional promiscuity gives the population a chance to be pre-adapted to the new conditions. As a consequence, evolutionary dynamics will be qualitatively different when promiscuity is pervasive. We devote the second half of the Chapter to the

study of quantitative models of evolutionary dynamics, studying empirical and simulated fitness landscapes where functional promiscuity is present.

4.1 Introduction

4.1.1 Functional promiscuity in molecular models

Despite their impressive specificity, enzymes are known to be very promiscuous (O'Brien and Herschlag, 1999; Khersonsky and Tawfik, 2010). Besides their main function, they are usually able to catalyze different reactions, even though they have not been selected for them (Aharoni et al., 2005; Amitai et al., 2007). This high promiscuity changes the view of cellular biology from a clockwork-like machine, with every pathway clearly connecting an input to an output, into a messy, sloppy system, in which spurious connections between different components are the norm rather than the exception (Daniels et al., 2008; Tawfik, 2010). This sloppy aspect of biology is no doubt a direct consequence of the "tinkering" character of evolution, as Jacob pointed out years ago (Jacob, 1977).

Functional promiscuity has important evolutionary consequences. Under selection for new functions, promiscuous enzymes can give selective advantage to the organism (Tokuriki and Tawfik, 2009; Tawfik, 2010), and represent starting points for the improvement of those functions, via gene duplication or other mechanisms (Aharoni et al., 2005; Amitai et al., 2007; Tokuriki and Tawfik, 2009; Khersonsky and Tawfik, 2010). Additionally, the exploration of new secondary —promiscuous— functions can be achieved through "neutral" mutations that do not alter the main function of the enzyme. The result is a heterogeneous population of molecules with hidden secondary functions, that become manifest when the environmental conditions change —what some have called cryptic genetic variation (Paaby and Rockman, 2014).

Promiscuity is not a property of enzymes alone. Other molecular systems, like RNA, GRNs and metabolic systems have been shown to express functional promiscuity —sometimes under different names, such as phenotypic plasticity (Wagner, 2011), latent phenotypes (Payne and Wagner, 2014) or exaptations (Barve and Wagner, 2013). In all cases, these systems

show the ability to perform more than one function. We will focus in this Chapter on RNA, Boolean GRNs and the metabolic phenotype of tOYLIFE.

As mentioned in Chapter 1, RNA sequences fold into three-dimensional structures called tertiary structures, that enable these molecules to perform their function. This tertiary structure is heavily conditioned by a two-dimensional one, the secondary structure, which is usually considered a good proxy for RNA function (Huynen, 1996; Aguirre et al., 2011). The folding into the secondary structure follows thermodynamical principles and biochemical constraints. Thermodynamics, in particular, guarantees that a given sequence will spend most of its time folded in the secondary structure that minimizes its free energy, and it is this one that is usually considered in genotype-phenotype map studies (Schuster et al., 1994; Fontana and Schuster, 1998; Jörg et al., 2008; Aguirre et al., 2011; Dingle et al., 2015). However, all other compatible structures —and their corresponding cellular functions— may be visited by the RNA molecule at some point during its lifetime, albeit with decreasing probability as the free energy associated to the structure increases (Ancel and Fontana, 2000; Wagner, 2014).

GRNs are able to develop several patterns of gene expression depending on the initial conditions (Espinosa-Soto et al., 2011; Payne and Wagner, 2014). GRNs can be modelled in various ways (Ciliberti et al., 2007a; Wagner, 2011; Payne et al., 2014), but we will focus here on the Boolean discrete model used by Payne et al. (2014), based on Kauffman's original work (Kauffman, 1969) —this is the same model that underlies tOYLIFE's regulation dynamics. Remember that in this model, genes can either be ON (1) or OFF (0), and the dynamics happens in discrete time. In a GRN formed by g genes, there are 2^g expression states: each of them represents one particular activation state for each gene. The genotype then maps every expression state —the input— to its corresponding output state. Thus, the expression state of the network at time t will determine which genes become expressed at time $t+1$. Because the genotype has to specify the activation state of each of the g genes in the network for each of the 2^g input states, it can be described as a binary string of length $L = g2^g$. Given any initial state, the regulatory dynamics of the network will reach an equilibrium, either a fixed-point or a periodic cycle —we call these attractors, and they are usually taken as the phenotype of the GRN. If a given GRN has

more than one attractor, this means that it can act differently —develop different gene expression patterns— depending on the initial state of the network. Because environmental cues can alter the expression state of any gene, they can take the network out of an attractor and into another. This is also functional promiscuity.

As for tOYLIFE, we will focus here on the metabolic phenotype studied in Chapter 3.

4.1.2 Functional promiscuity as a multiplex network of genotypes

The general picture of functional promiscuity suggests an abstract extension of the neutral network framework, which we have already discussed throughout this thesis (see Chapter 1). In this framework, genotypes are taken to express just one phenotype (Figure 4.1**a**). Promiscuous genotypes, however, have more than one phenotype. We can understand that two genotypes expressing different promiscuous phenotypes belong to different neutral networks, as in Figure 4.1**a**. However, this does not reflect the complexities of the evolutionary process. At a given time, a limited set of functions will be selected for, and those genotypes expressing the same phenotype with respect to that function will be part of the same neutral network. But they could belong to different neutral networks in a different environment or under different selective pressures. The set of neutral networks will have to be considered separately for each environment.

This situation is well grasped by a general model recently introduced in the literature: a multiplex network (Kivelä et al., 2014). A multiplex network is a network made out of different 'layers'. The same nodes appear in each layer, but their connections can be different. The dynamics going on on the network takes place in each layer separately, but given that the same node belongs to different layers, what happens in one of them will have some effect on the other layers.

In our case a layer corresponds to a given environment. Using a metabolic system, such as tOYLIFE, as a way of example, each environment will be represented by the presence or absence of a given metabolite. If a node, corresponding to a genotype, can catabolize it, then it will be viable in that environment. Each layer contains the neutral network associated to its environment —formed by all viable nodes in that environment—, so the

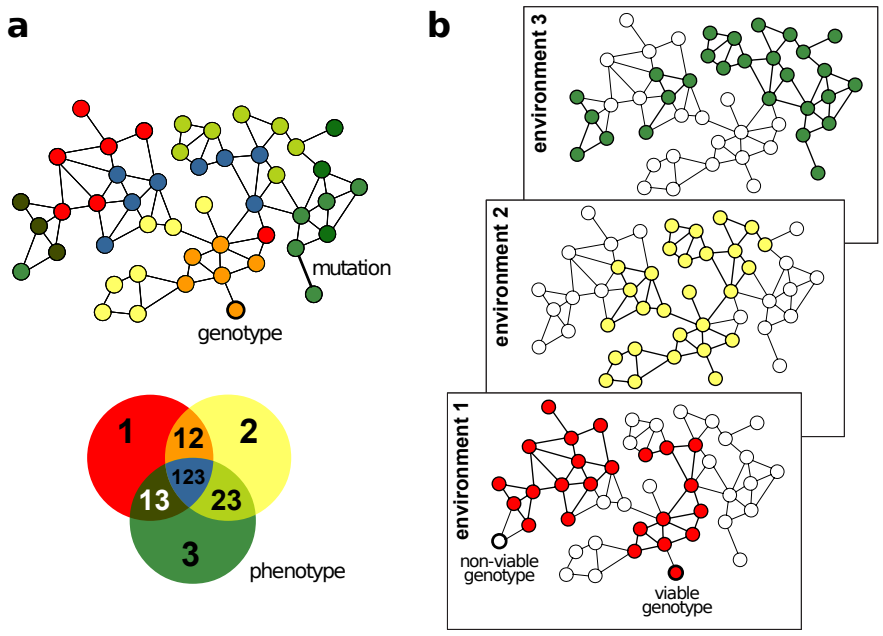


Figure 4.1: Sketch of a neutral network and a multiplex network of genotypes. (a) We can assign one color to each genotype according to its phenotypes. Genotypes showing phenotype 1 are colored in red, and those expressing both phenotypes 1 and 2 are colored in orange. Traditional neutral networks are not good intuitive tools to grasp the presence of promiscuity: we do not know in which conditions each genotype will express which phenotype, and therefore we cannot make any assumptions about neutrality. **(b)** A better representation of this situation is achieved through a multiplex network. Multiplex networks are made of several layers with the same nodes —although the links between them can be different—, one for each different environment. In the figure, each of the three layers corresponds to a different environment, selecting for each phenotype in **a**. Colored circles represent nodes that are viable in that environment (according to the code defined in **a**), whereas empty circles represent non-viable nodes —non-viable nodes act as if they were not present in the network. This is a very simple fitness landscape, but the multiplex can easily accommodate more complex ones. Note that the neutral networks in the three layers are all different, although some nodes appear in two or even the three of them (these common nodes are marked with a different color in **a**).

multiplex network can be regarded as a stack of different neutral networks for the same set of nodes (Figure 4.1**b**). Interestingly, networks can have several disconnected clusters in every layer although the network of all viable nodes —in at least one environment— forms a single connected cluster. This provides a connectivity between unreachable genotypes through environmental changes. This feature assigns to the environment an unexpected role as a facilitator of adaptation, as has already been pointed out in De Vos et al. (2015) and Steinberg and Ostermeier (2016). This simplified picture can easily be extended to accommodate for more complex fitness landscapes: we could include more than one neutral network per layer, fitness differences, and so on —and, indeed, we will do this later on in this chapter (see Section 4.3).

The multiplex framework is general and not only valid for $\mathtt{toy}$LIFE. In GRNs, each layer represents a given initial state. For each of these, a genotype will express one particular attractor, shared with other genotypes. For RNA, the picture becomes less clear, but we can interpret each layer as an environment where a given secondary structure is selected for. In this case, different genotypes expressing the same structure could have different fitness values due to differences in folding energy. Besides its mathematical applications, we believe the multiplex to be a powerful tool to help us in our intuitive understanding of evolution.

In this Chapter, we want to explore the prevalence and significance of promiscuity in RNA, Boolean GRNs and $\mathtt{toy}$LIFE. We will study how new functions are discovered through promiscuity. Finally, we will study evolutionary dynamics on a multiplex network in a shifting environment, showing some unexpected outcomes with possible applications.

4.2 Prevalence of promiscuity in genotype-phenotype maps

4.2.1 Measures and prevalence of promiscuity

We have used different sensible measures of functional promiscuity for each of the three genotype-phenotype models. Each of these measures depends on the specific definition of phenotype and environment for each model. For the RNA genotype-phenotype map, the measure of promiscuity is the number of secondary structures a single sequence can fold into,

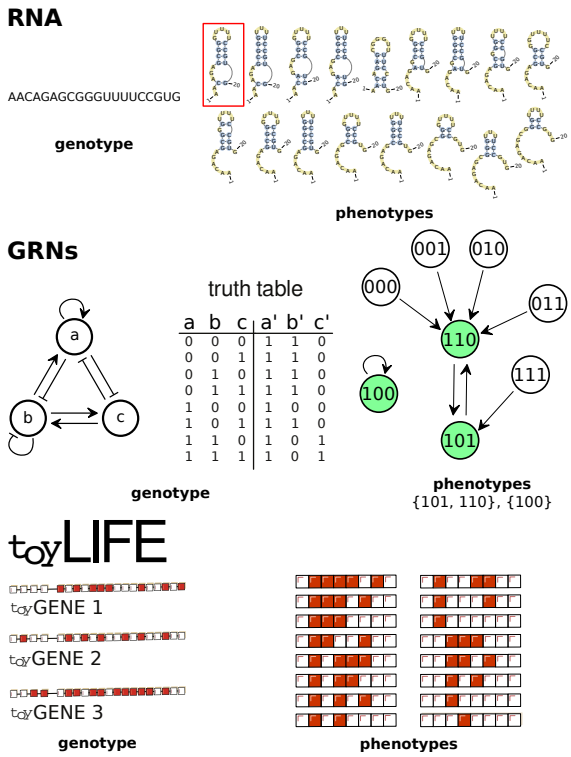


Figure 4.2: Genotype-phenotype maps and definition of promiscuity. (Upper row) RNA genotypes (upper row) are sequences that fold into various secondary structures, or phenotypes. The minimum free energy structure is outlined in red. This sequence folds into 17 different structures, and so its promiscuity is 17. Figures of secondary structures were obtained with the PseudoViewer web service (Byun and Han, 2006, 2009). **(Middle row)** GRN genotypes are binary strings of length $g2^g$, where g is the number of genes in the network, specifying the expression state of each gene for each input state. The phenotype is the expression dynamics of this network, which in this case consists of two attractors: one formed by the state 100, and the other formed by the cycle 101-110. The rest of initial states map onto this second attractor. The promiscuity of this network is 2. **(Lower row)** toyLIFE genotypes are sets of genes. The phenotype, computed following the rules defined in Chapter 2, is the set of metabolites this genotype is able to catabolize. In this case, this genotype is able to catabolize 16 different metabolites, and therefore its promiscuity is 16.

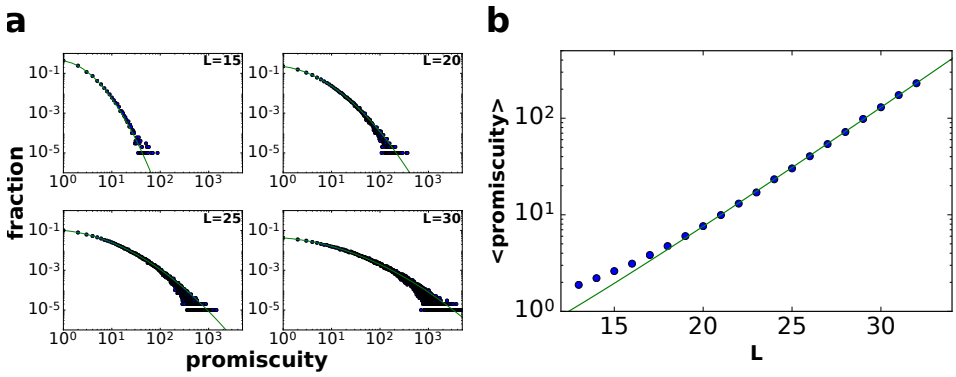


Figure 4.3: Functional promiscuity in RNA. (a) Probability distributions for functional promiscuity in RNA, for sequences of length $L = 15, 20, 25$ and 30 (blue circles). We sampled $100,000$ sequences for each L. All four distributions can be fitted to a parabola in this log-log plot (green lines), suggesting that these distributions are truncated log-normals. The fit was done using the least squares method. **(b)** Average promiscuity scales with L as $\langle \text{promiscuity} \rangle = 0.125 L^{-0.718} 1.367^{L}$.

independent of their free energy —as long as it is lower than the energy associated with the open structure (Figure 4.2, upper row). In order to obtain all secondary structures compatible with a single sequence, we used the `subopt` routine from the Vienna 2.2.10 package (Wuchty et al., 1999; Lorenz et al., 2011), with default parameters.

We have used sequences from length $L = 13$ to length $L = 32$, randomly sampling $100,000$ sequences for each L, and measuring their functional promiscuity. The distribution of promiscuity in RNA follows a truncated log-normal distribution (Figure 4.3**a**). That is, the decimal logarithm of promiscuity follows a truncated normal distribution —a normal distribution bounded below. The truncation is natural, because the minimum value for promiscuity is 1. However, the parameters of the log-normal distribution before its truncation are not fixed, and depend on L. This peculiarity makes it harder to write the distribution in closed form. Besides, we have no insight into what might be behind these distributions of promiscuity, which will surely depend on the particularities of the folding process in RNA. In Figure 4.3**a** we have fitted the logarithm of the data to parabolas (solid lines) to highlight the proximity to this truncated log-normal distribution. It is remarkable, however, that the number of sequences that fold

into a given structure —the dual of this distribution— also follows a log-normal law (see Chapter 1 and Dingle et al. (2015)).

Many RNA sequences fold only in one structure, but the probability that a sequence folds into many structures is not negligible. The average promiscuity grows almost exponentially with L: $\langle \text{promiscuity} \rangle = 0.12L^{-0.72}1.4^L$ (Figure 4.3**b**). Therefore, the average promiscuity for a sequence of $L = 20$ is below 10, but for $L = 30$ it is over 100. The total number of structures in phenotype space grows as $1.5L^{-1.5}1.8^L$ (Schuster et al., 1994; Aguirre et al., 2011), which means that the relative fraction of structures per sequence goes to zero as L increases. In other words, although longer RNA sequences are exponentially more promiscuous, the fraction of possible structures they can explore is negligible. However, we argue that the relevant number here is the absolute number of promiscuous structures. The near exponential growth of this number shows that RNA molecules tend to become more and more promiscuous as they grow in length, thus favoring their ability to find new functions.

In a similar analysis, Wagner found that the average number of promiscuous phenotypes also increases exponentially (Wagner, 2014). His definition of promiscuity, however, is slightly different from ours. We have considered all secondary structures compatible with a given sequence, as long as the folding energy was below 0 kcal/mol. Wagner considered structures with a folding energy within a certain range above the minimum free energy. His results are qualitatively identical to ours. We have repeated our simulations with an alternative definition of promiscuity, closer to Wagner's, and our main conclusions remain unaltered (see Figure 4.4).

For GRNs, the measure of promiscuity must be different. Cells are constantly exposed to inputs from the changing environment. Proteins, sugars, small peptides, antibiotics, toxins... many different molecules contact the cell membrane at a given moment (Phillips et al., 2012). As a consequence, they need to be prepared to change their regulatory expression pattern to accommodate the changes in the environment. Different regulatory programs for different environmental challenges are a given in cell biology (Alberts et al., 2014), and they are an expression of functional promiscuity at the regulatory level. Our definition of promiscuity must reflect these particularities, so we have used the number of different attractors associated to each network, without taking into account which initial states lead

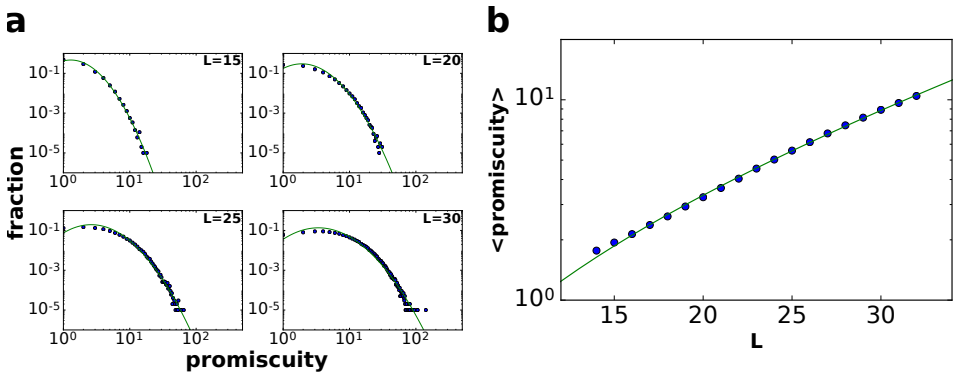


Figure 4.4: Alternative promiscuity in RNA. We can consider an alternative definition of promiscuity for RNA, similar to the one considered in Wagner (2014). In order for a structure to belong to the promiscuous repertoire of a sequence, we will only consider those whose free energy is no more than 200 kcal/mol higher than the energy associated to the minimum free energy structure. Our results hold qualitatively, although the scaling of the log-normal parameters with L changes. **(a)** Probability distributions for functional promiscuity in RNA, for sequences of length $L = 15, 20, 25$ and 30 (blue circles). We sampled $100,000$ sequences for each L. All four distributions can be fitted to a parabola in this log-log plot (green lines), suggesting that these distributions are truncated log-normals. The fit was done using the least squares method. Note, however, that the fit is not so good as in Figure 4.3. **(b)** Average promiscuity scales with L as $\langle \text{promiscuity} \rangle = 0.059 L^{0.956} 1.060^{L}$.

to them (Figure 4.2, middle row). More promiscuous GRNs will yield a larger number of regulatory outputs when presented with different input states.

As with RNA, we have randomly sampled a large number of genotypes and measured their promiscuity. We used networks formed by $g = 3$ to $g = 15$ genes, sampling $100,000$ networks for $g \leq 10$ and $10,000$ networks for $g > 10$. The distribution of promiscuity follows, asymptotically, a shifted Poisson distribution —shifted because it starts at 1, the minimum value for promiscuity, instead of 0, like usual Poisson distributions do (Figure 4.5**a**). Techniques from analytic combinatorics (Flajolet and Sedgewick, 2009) allow us to obtain this asymptotic fit and derive an expression for the parameter of the distribution: $\lambda = (\log 2)(g-1)/2$ (see Flajolet and Sedgewick (2009, p.449) and Appendix A.1). The fit between

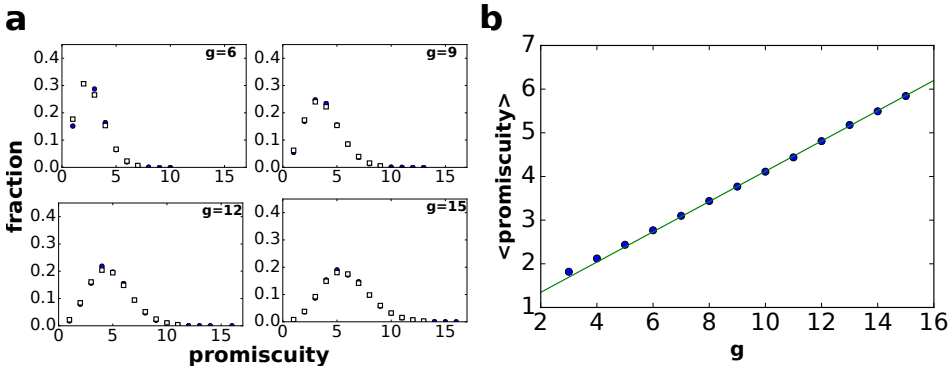


Figure 4.5: Functional promiscuity in GRNs. **(a)** Probability distributions for functional promiscuity in GRNs, for networks formed by $g = 6, 9, 12,$ and 15 genes (blue circles). We sampled $100,000$ networks for $g = 6$ and $g = 9$, and $10,000$ for $g = 12$ and $g = 15$. Using techniques from analytic combinatorics (see text), we found an asymptotic fit for the distribution, a shifted Poisson with parameter $(\log 2)(g-1)/2$ (white squares). **(b)** Sample averages (circles) and theoretical prediction of the average, which is $1 + (\log 2)(g-1)/2$ (solid line).

the theoretical prediction and the experimental data is very good, even for small network sizes (Figure 4.5**b**). This means that the average number of functions grows linearly with gene number. As in RNA, the number of GRN phenotypes grows much faster than the average promiscuity, as $(2^g - 1)!$ (see Appendix A.2). This means that each genotype will explore a vanishingly small fraction of the set of phenotypes. As with RNA, however, we argue that the relevant quantity is the absolute promiscuity, as it reflects the ever-growing ability of genotypes to generate promiscuous functions. Our simulations only explore networks with a reduced number of genes. Extrapolating the average number of functions, however, we obtain that for a network formed by $4,000$ genes, such as *E.coli*, the average number of functions it can express is over $1,000$. Of course, real regulatory networks need not be similar to those found at random in genotype space, but this high number suggests that the ability of real regulatory networks to develop different regulatory functions is not especially constrained, and that cells can easily manage the number of environments they will find in their lifetime.

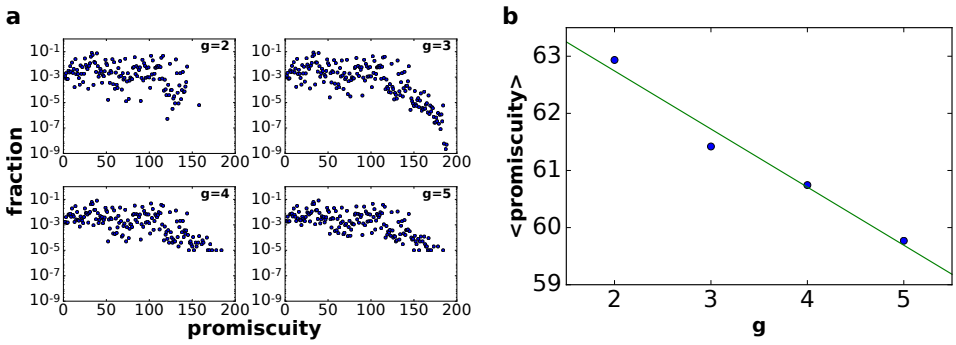


Figure 4.6: Functional promiscuity in tQYLIFE. (a) Probability distributions for functional promiscuity in tQYLIFE, for genotypes formed by $g = 2, 3, 4,$ and 5 genes (blue circles). **(b)** The average promiscuity decreases linearly with gene size, as $\langle \text{promiscuity} \rangle = 64.8 - 1.02g$.

Finally, for tQYLIFE, promiscuity is defined as the number of metabolites that a genotype is able to catabolize (Figure 4.2, lower row). This measure is similar to the one used by Barve and Wagner (2013) for a different metabolic model. We have studied genotypes formed by $g = 2$ to $g = 5$ genes, sampling $100,000$ genotypes for $g = 4$ and $g = 5$ and using the exhaustive data obtained in Chapter 3 for $g = 2$ and $g = 3$.

Promiscuity in tQYLIFE is very high even for small genotypes (Figure 4.6**a**): genomes containing as few as two genes are already able to metabolize, on average, 62.9 metabolites out of 214 (see Chapter 3) —note that, in this case, the number of possible phenotypes remains constant, as opposed to what happened in our previous examples. Interestingly, this average decreases linearly as genotype size increases (Figure 4.6**b**): $\langle \text{promiscuity} \rangle = 64.8 - 1.02g$, at least for small g. If we combine this decrease with the decreasing influence of the phenotypes that appear in $g = 2$ —that we saw in Chapter 3—, these results suggest that the phenotypes in $\mathcal{P}_2$ are more promiscuous than the phenotypes produced by more complex genotypes, in which more than two genes are influencing metabolism. Although far from a realistic depiction of metabolism, this result suggests interesting features of complex metabolism, in which the interaction of many genes generates constraints on the output.

4.2.2 Discovery of new phenotypes through promiscuity

For each map, we randomly selected $100,000$ genotypes and performed a neutral random walk on individual layers of the multiplex neutral network —the latter was defined in section 4.1.2. The neutral network in each layer is formed by (a) all genotypes that express the same minimum free energy structure in the case of RNA, (b) all genotypes that express one particular attractor in the case of GRNs, and (c) all genotypes that are able to metabolize one particular metabolite in toyLIFE. Every time step, we attempted a point mutation at a randomly chosen genome site, by changing one letter of the sequence for another one chosen at random from the alphabet — $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$ in RNA and $\{0,1\}$ in GRNs and toyLIFE. If the new genotype belonged to the original layer, the mutation was accepted. Otherwise, the mutation was discarded and the process was repeated until a neutral mutation was obtained. We repeated this process $T = 1,000$ times, for each seed. For RNA, we used genotypes of length $L = 20$. For GRNs, we used networks made of $g = 5$ genes. Finally, for toyLIFE, we used genotypes formed by $g = 3$ genes.

Each time a mutation occurs, the set of secondary phenotypes of the new genotype can change. We compared these sets, recording the cumulative number of new phenotypes "discovered" in each random walk, as well as the difference in secondary phenotypes between time step t and $t-1$ —we term the latter quantity $\Delta\text{promiscuity}(t)$. The results are shown in Figure 4.7. The left panels show the average increase of cumulative phenotypes discovered by promiscuity, whereas the right panels show the distribution of $\langle\Delta\text{promiscuity}\rangle = \frac{1}{T}\sum_{t=1}^{T}\Delta\text{promiscuity}(t)$ for all random walks.

The discovery of new phenotypes through promiscuity follows very different dynamics in the three models. In RNA, the probability that a change in the set of secondary phenotypes occurs is, on average, 0.36 — that is, roughly, one change every 3 time steps. The magnitude of change is, on average, 3.4: when the set of secondary phenotypes does change, on average 3 new phenotypes are discovered. As a result, the overall average magnitude of change per time step is 1.4 new phenotypes. However, some phenotypes are re-discovered more than once along a random walk, and therefore the cumulative number of secondary phenotypes could saturate at some point: Figure 4.7 shows that this is not the case: although on

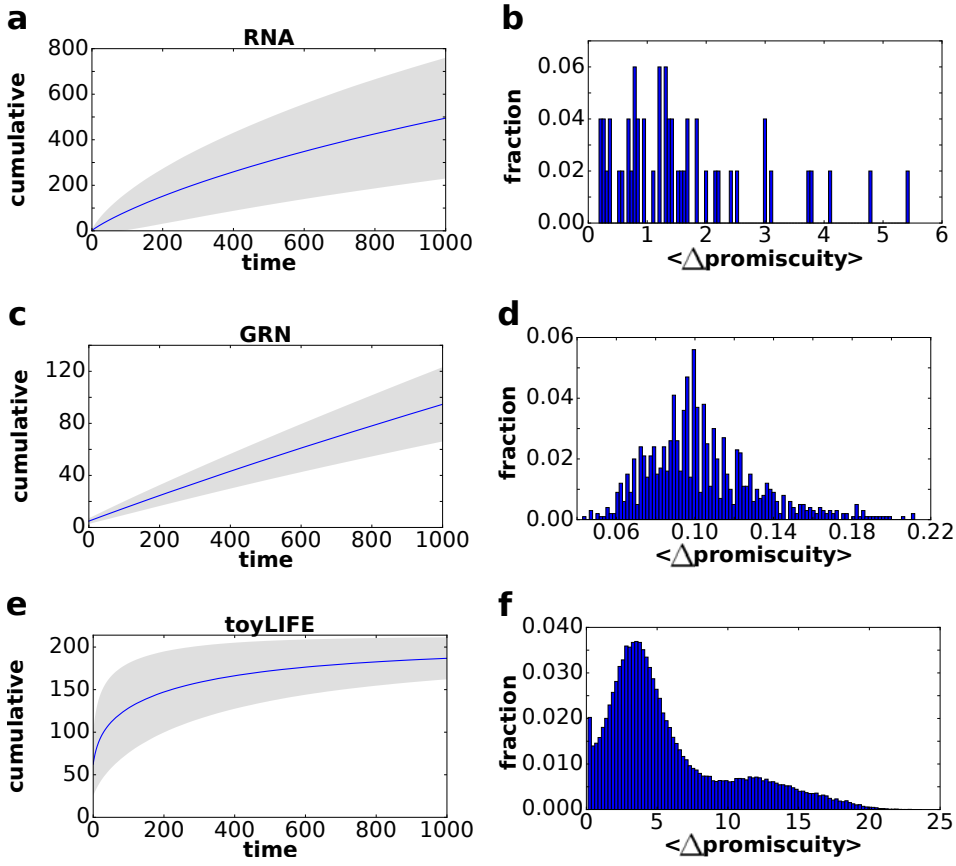


Figure 4.7: Discovery of new phenotypes in a neutral random walk. (Left) Cumulative promiscuity along a neutral random walk, in RNA (**a**, $L = 20$), GRNs (**c**, $g = 5$) and toyLIFE (**e**, $g = 3$). The plots show the average increase in cumulative promiscuity (blue lines) plus minus one standard deviation (gray area), computed from $100,000$ random walks. **(Right)** Distribution of promiscuity changes along a neutral random walk, for the same random walks in RNA (**b**), GRNs (**d**) and toyLIFE (**f**). For each random walk, we measured the instantaneous change in promiscuity from one time step to the next, Δpromiscuity (see text). The figures show the distribution of $\langle \Delta\text{promiscuity} \rangle = \frac{1}{T} \sum_{t=1}^{T} \Delta\text{promiscuity}(t)$ for the $100,000$ neutral random walks, where T is the length of the random walk.

average the increase in the cumulative number of secondary phenotypes slows down at the beginning, it keeps growing linearly up until the first

1,000 mutations. This almost linear growth of the cumulative number of discovered phenotypes has also been observed by Wagner (2014) with his alternative definition of promiscuity. Ancel and Fontana (2000) found that sequences that have a particular structure as a promiscuous phenotype are close in genotype space to sequences that express that same structure as the minimum free energy structure. This phenomenon has also been called "look-ahead effect" in more general terms (Whitehead et al., 2008). The look-ahead effect effectively increases the genotype's fitness, by making it express a fitter phenotype, albeit probabilistically. This would increase the promiscuous genotypes' frequency in the population until some mutation finds one genotype that expresses the fitter phenotype constitutively. In this light, promiscuous genotypes could function as a stepping-stone for the selection of new stable structures (Wagner, 2011). RNA molecules would therefore be another example of the genetic assimilation mechanism proposed by Waddington (1953).

In GRNs, the probability of discovery is 0.10, roughly one change in the set of phenotypes every ten time steps. Because of the way mutations are implemented in this model, the maximum number of new phenotypes discovered each time step is one —a mutation changes only one arrow in the attractor graph pictured in Figure 4.2, so either a new attractor is created or not. Thus, a new phenotype is discovered, on average, every ten time steps —this can be checked in Figure 4.7, which shows an almost perfectly linear increase. Adding this high potential to find new, unexplored functions to the high number of promiscuous functions any network can express, this result gives hints as to the ease with which GRNs can find new functions to respond to the environment. In a similar study with a different model for GRNs, Espinosa-Soto et al. (2011) found that the cumulative number of discovered phenotypes along a neutral random walk also increased linearly, without any hint of saturation. They also found that, akin to RNA, genotypes that present a given function secondarily are close in genotype space to those that generate it primarily. Again, this phenomenon facilitates adaptation through promiscuity.

Finally, for $\mathrm{t_{OY}}$LIFE the dynamics is slightly different. The average probability of change is 0.04, or one change every 25 time steps. However, being a metabolic phenotype, the average magnitude of change is much larger, around 15.2. In this case, there is no linear growth, and the satu-

ration in the cumulative number of discovered phenotypes is much more evident. Remember that the number of phenotypes is fixed and equal to 214, which explains this saturation. Observe, however (Figure 4.7) that the average cumulative number of secondary phenotypes is very close to the maximum, which means that toyLIFE genotypes can easily find all metabolic functions by promiscuity, a property that was already hinted at in Chapter 3.

For all three models, the ability to discover new phenotypes through promiscuity is not negligible. In the presence of environmental changes, different members of the same population will express different promiscuous phenotypes, making it easier for the population to adapt to the new conditions. We will explore the dynamical consequences of the presence of promiscuity in the next section.

4.3 Shifting environment dynamics

We studied the dynamics of shifting —alternating— environmental change on three small multiplex genotype networks (see Figure 4.8) that show some degree of functional promiscuity: there are genotypes in those networks that are able to survive in both environments. Our interest lies in studying how the frequency of environmental change affects population structure and number. The ulterior motivation for these questions is to understand the conditions needed to eliminate a population through controlled environmental change —this topic has important medical applications that will be discussed later on.

The three networks are qualitatively different. The first one (Figure 4.8**a**) is a synthetic network (we will refer to this network as N12 from now on). It contains 12 nodes, and only two values of fitness, 1 and 0. That is, a genotype is either viable in the environment or not. The fitness values have been carefully chosen so that the set of genotypes that is viable in each layer is different —but there is still some overlap between the layers, indicating some degree of functional promiscuity. The values were also chosen so that both layers are completely uncorrelated (Figure 4.9**a**). The second network (Figure 4.8**b**) is an experimental fitness landscape taken from De Vos et al. (2015), and the fitness values have been kindly shared by Marjon de Vos (we will refer to this network as N64 from now on). This

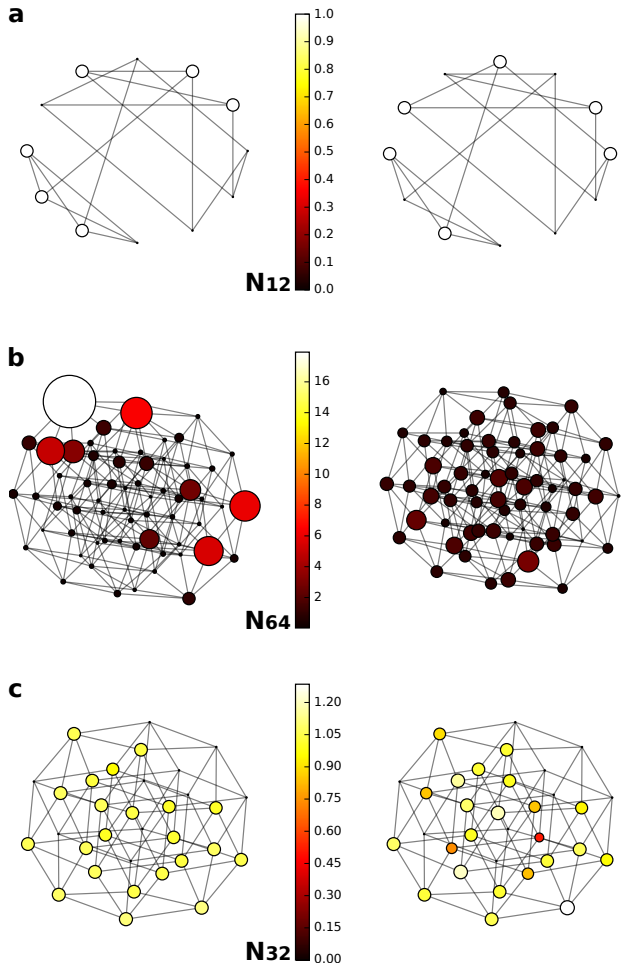


Figure 4.8: Example multiplex networks. The three example networks we have used for our dynamical simulations. Nodes' fitness values are represented through a color code (the values are in the color bar), and the node's radius is also proportional to its fitness, for clarification —nodes with null fitness appear as little more than dots. Edges are drawn if two genotypes are connected by point mutations. **(a)** Synthetic network (N12), with 12 nodes and only two values of fitness —1 and 0. Genotypes are either viable or non-viable in a given environment. **(b)** De Vos et al. (2015)'s landscape (N64), kindly shared by Marjon de Vos. There are 64 nodes. **(c)** Cervera et al. (2016)'s landscape (N32), kindly shared by Santiago F. Elena. There are 32 nodes.

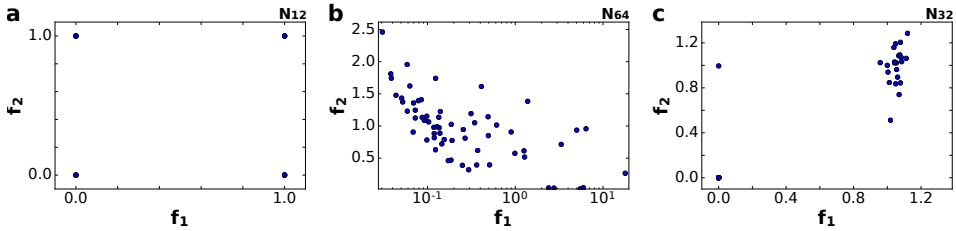


Figure 4.9: Example multiplex networks. (a) Fitness values in the first ($\mathbf{f}_1$) and second ($\mathbf{f}_1$) environment for N12. The correlation coefficient $\rho = 0$. We have every combination of viable / non-viable possible. **(b)** Fitness values for N64. There is a negative correlation between $\mathbf{f}_1$ and $\mathbf{f}_2$, $\rho = -0.33$. That means that genotypes that are very fit in one environment will be very unfit in the other. **(c)** Fitness values for N32. The correlation is positive and high, $\rho = 0.89$, with just one genotype being non-viable in one environment and with an average fitness in the other.

network reflects the repression (Figure 4.8**b**, left) and expression (Figure 4.8**b**, right) ability of *lac*-repressor-operator *E. coli* mutants. In the absence of lactose, cells are fitter if they are better at repressing the expression of the operon —an operon is a collection of genes that are expressed jointly. In the presence of lactose, bacteria are fitter if they are able to better express the operon. Thus, the presence of lactose defines the environmental change. De Vos et al. (2015) studied the repression and expression ability of 64 mutants —the network is a hypercube of dimension 6. The network is particular in that fitness values are negatively correlated between both environments: genotypes that are very fit in one layer are maladapted to the other layer, and vice versa (Figure 4.9**b**). Finally, the last network (Figure 4.8**c**) is another experimental fitness landscape taken from Cervera et al. (2016) —the fitness values were kindly shared by Santiago F. Elena (we will refer to this network as N32 from now on). This network reflects the growth rates of 32 genotypes of TEV virus in two different hosts, *Arabidopsis thaliana* (Figure 4.8**c**, left) and *Nicotiana tabacum* (Figure 4.8**c**, right). The environmental change is thus defined as a change of host. This network is a hypercube of dimension 5, with 32 nodes. In this case, the fitness values are positively correlated, with a correlation coefficient of $\rho = 0.89$ (Figure 4.9**c**). Fitness values in N12 and N32 are around 1, with non-viable nodes having fitness 0. For N64, the fitness values in De Vos

et al. (2015) have been normalized so that the average value is 1. However, they are much more diverse than the other two networks (Figure 4.8).

In order to study how the composition of the population changes as a function of the frequency of environmental change, we studied a discrete-time, infinite-population model based on quasi-species dynamics (Aguirre et al., 2009). The dynamical equations of the model are:

$$x_i(t+1) = \frac{1}{\phi(t)}\left((1-\mu)f_i(t)x_i(t) + \frac{\mu}{k_{\max}}\sum_{j=1}^{G}\mathbf{A}_{ij}f_j(t)x_j(t)\right), \quad i = 1,\ldots,G.$$
$$\phi(t) = \sum_{i=1}^{G} f_i(t)x_i(t), \tag{4.1}$$

where $x_i(t)$ is the fraction of the population that occupies node (genotype) i at time t, μ is the mutation rate, $f_i(t)$ is the fitness of node i at time t, which will depend on the environment that the population is in at time t (each environment's fitness values are given in Figure 4.8), $k_{\max}$ is the number of neighbors per genotype, G is the size of the network and $\mathbf{A}$ is the adjacency matrix of the network: $\mathbf{A}_{ij}$ is 1 if nodes i and j are neighbors, and 0 otherwise. This system of equations means that the fraction of the (infinite) population that occupies node i at time $t+1$ will have a contribution from all the un-mutated offspring of node i at time t, plus all the mutants descended from its neighbors. The right-hand side of the equation is divided by $\phi(t)$, the average fitness of the population at time t, for normalization. We will assume $\phi(t) > 1$, so that the population does not become extinct —in that sense, we take the values of $f_i(t)$ to be relative terms, relevant in order to compute the composition of the population, but independent of population growth.

The change of environment is only reflected in equations (4.1) in the $f_i(t)$ values. Each node i has associated two values of fitness, $f_i^{(1)}$ and $f_i^{(2)}$ —its fitness in the first and second environments, respectively. The parameter that modulates the frequency of environmental change is p. Starting at one of the environments, the environment changes with a probability p, akin to an average frequency of change. When $p = 1$, the environment changes every time step. When $p < 1$, the population stays in a given environment for g generations, where g is a random number chosen from a ge-

ometric distribution with parameter p, and then the environment changes. When $p \to 0$, the population stays in each environment an infinite amount of time.

The average occupation of each node, given by

$$\bar{x}_i(t) = \frac{1}{t} \sum_{\tau=1}^{t} x_i(\tau), \tag{4.2}$$

reaches an equilibrium value $\bar{x}_i^*$ when $t \to \infty$: the environment keeps changing, but the average occupation of each node remains constant. This value of $\bar{x}_i^*$ depends on p, and we computed the equilibrium vector $\bar{\mathbf{x}}^*(p)$ for different values of p from 0 to 1. Because we want to study the effect that p has on the composition of the population, we will define a measure of distance between these vectors, in order to better visualize the results. This measure is the cosine distance, defined as

$$S(\mathbf{x}, \mathbf{y}) = \frac{\sum_i x_i y_i}{(\sum_i x_i^2)^{1/2} (\sum_i y_i^2)^{1/2}}, \tag{4.3}$$

where x_i is the i-th component of vector $\mathbf{x}$. Note that the numerator is the standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ in $\mathbb{R}^d$, and that the factors in the denominator are the standard Euclidean norms $\|\mathbf{x}\|_2$ and $\|\mathbf{y}\|_2$. Thus, S is the cosine of the angle formed by vectors $\mathbf{x}$ and $\mathbf{y}$, taking values from 0 —when $\mathbf{x}$ and $\mathbf{y}$ are most dissimilar— to 1 —when they are equal. For our purposes, we will use

$$S(p) = S(\bar{\mathbf{x}}^*(p), \bar{\mathbf{x}}^*(1)), \tag{4.4}$$

that is, we will always compare the equilibrium composition of the population obtained for any p with the one obtained when $p = 1$.

The equilibrium composition changes visibly as a function of p, as we can see in Figures 4.10, 4.11 and 4.12. When $p = 1$, the population is concentrated on the nodes that are fitter in both environments. The degree of concentration depends on μ: higher values of μ are associated with a more diverse population. For the N12 network, that means those nodes for which $f_i^{(1)} = f_i^{(2)} = 1$ (Figure 4.10). For the N64 network, because fitness values are anti-correlated, the population is concentrated on a few nodes for which fitness is not very low in both environments (Figure 4.11). For the

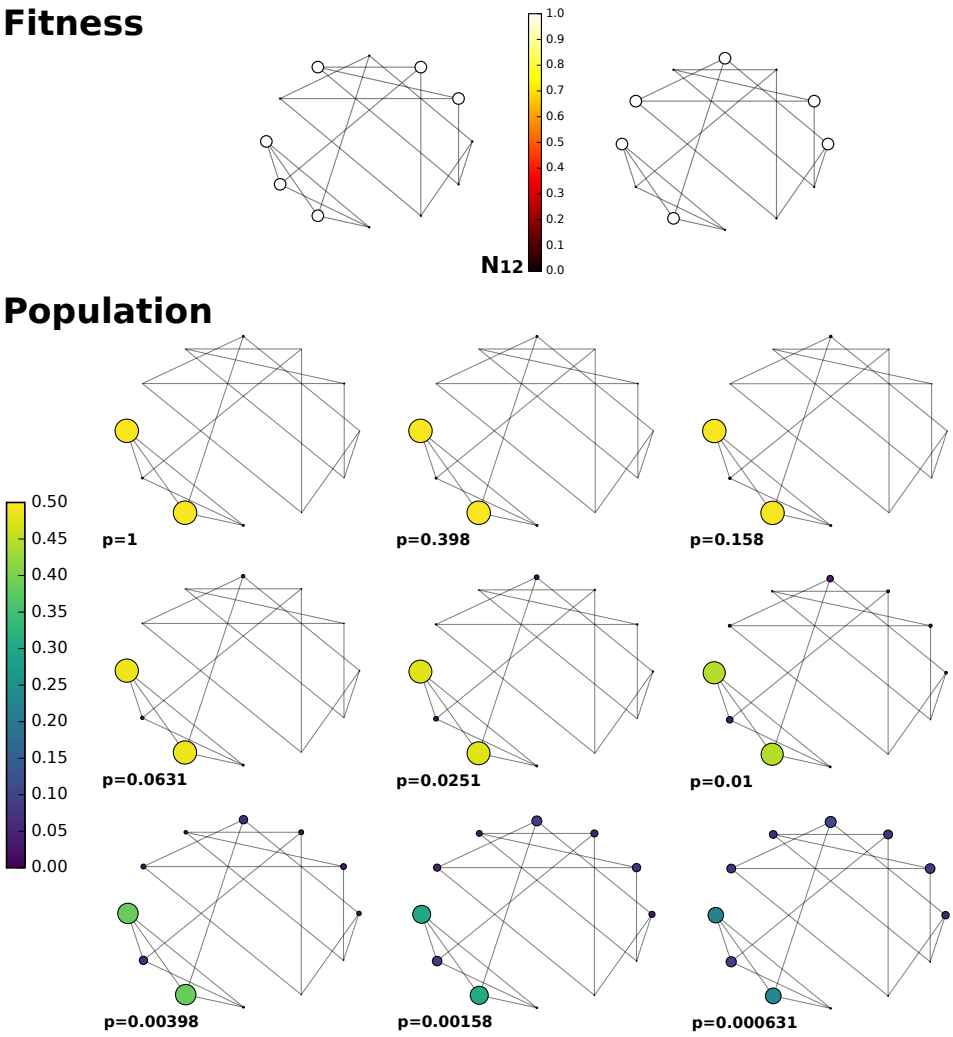


Figure 4.10: Average equilibrium occupation of each node as a function of p (N12). (Upper half) The fitness values associated with both environments for the N12 network, as in Figure 4.8. They are shown again here for reference. (Lower half) The equilibrium composition of the population is represented through a color code: violet for low occupation, yellow for high occupation. The nodes' radii are proportional to their average occupation, for clarification. Edges are drawn if two genotypes are connected by point mutations. For these simulations, $\mu = 0.01$.

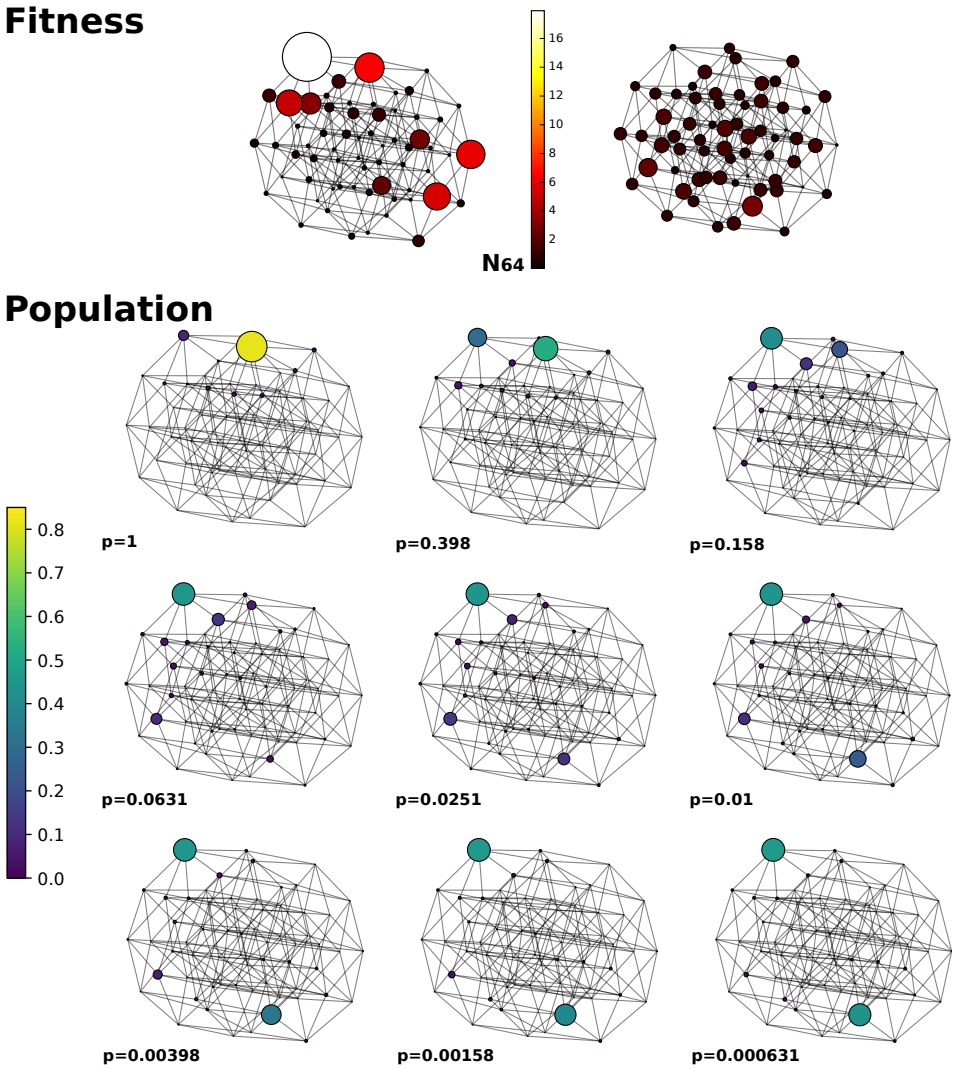


Figure 4.11: Average equilibrium occupation of each node as a function of p (N64). Same as Figure 4.10, but for N64.

N32 network, the population is concentrated around the fitter nodes, because they are fit in both environments (Figure 4.12). As p decreases, the equilibrium shifts, and as $p \to 0$ the equilibrium looks like an average between the equilibrium compositions of both environment taken separately.

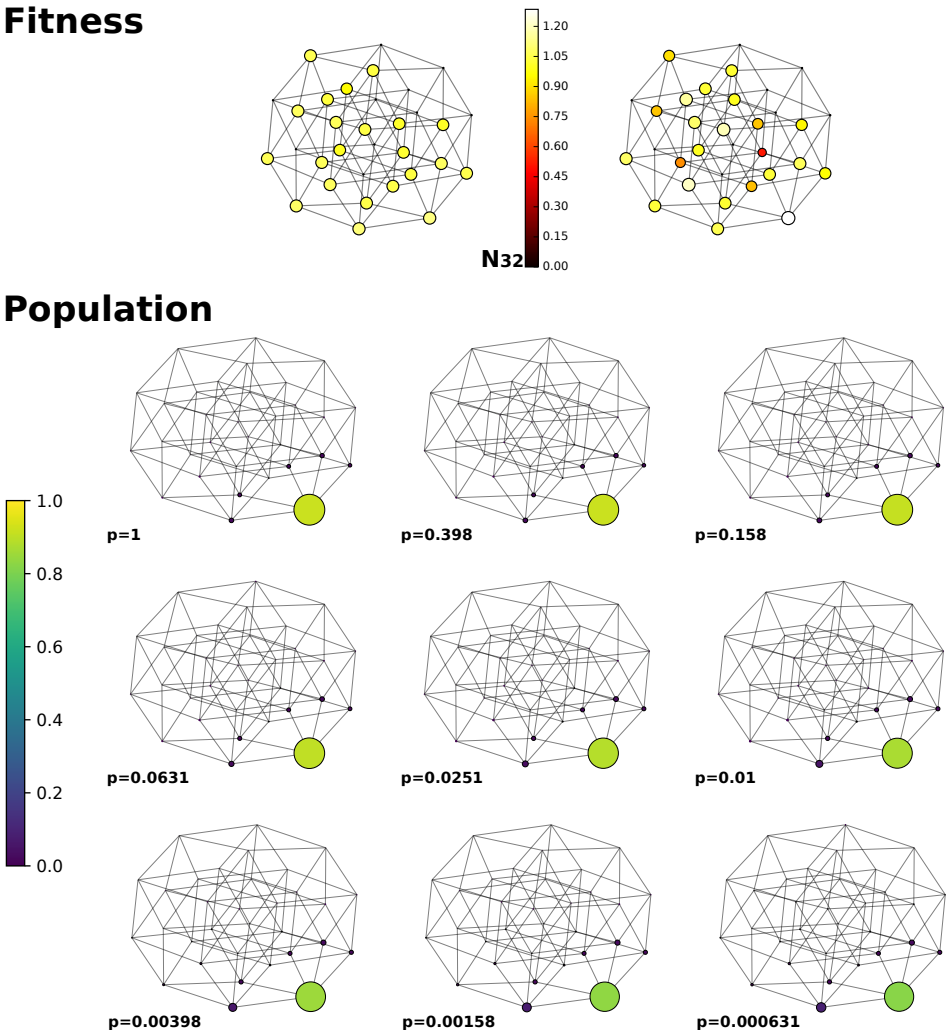
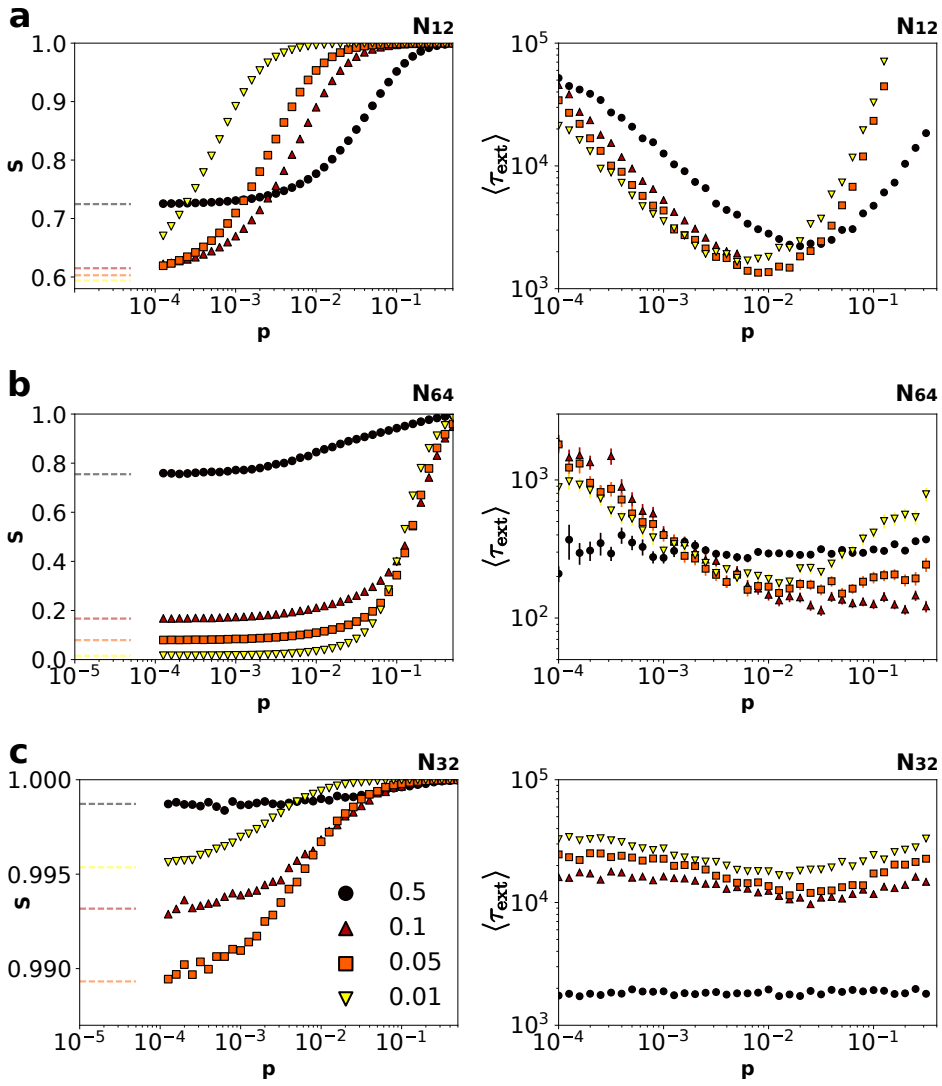


Figure 4.12: Average equilibrium occupation of each node as a function of p (N32). Same as Figure 4.10, but for N64.

Thus, for N12, nodes that are viable in one environment but not in the other will be populated to some extent. For N64, the population will be dominated by the two nodes that are fitter in each environment. Finally, for N32 the composition remains very similar. The similarity between com-



positions, as measured by S, can be observed in Figure 4.13 (left panels). In all three networks, the equilibrium changes smoothly from the $p = 1$ case to the $p \to 0$ limit, following a sigmoidal curve. Note, however, that the values of S span very different ranges in all cases. In N12, S is never smaller than 0.5, whereas for N64 it can get as low as 0.1. For N32, the two environments are so similar that S is always very high.

Figure 4.13: (Previous page.) **Evolutionary dynamics in a shifting environment.** Results for the simulations of our infinite and finite population models for **(a)** N12, **(b)** N64 and **(c)** N32, for different mutation rates: $\mu = 0.5$ (black circles), 0.1 (red triangles), 0.05 (orange squares) and 0.01 (yellow triangles). The left panels show the equilibrium composition of the population —visualized here with variable S— as a function of p, the frequency of environmental change, in the infinite population model. The similarity with the case $p = 1$ decays in a sigmoidal fashion, with different speed in the three networks. The dashed lines show the value of S in the limit $p \to 0$. Note how this value is approximated by our simulations, even though the values of p are still high. The right panels show average time to extinction $\langle \tau_{\text{ext}} \rangle$ as a function of p —each point is an average of 500 realizations. Note that there is (almost) always a critical value of p for which $\langle \tau_{\text{ext}} \rangle$ is minimized. Note that the y-axis is in logarithmic scale. Markers represent sample means, and error bars represent the standard error of the mean. See text for more details. For these simulations, $K = 50$ (for N12 and N64) and $K = 25$ (for N32).

In all cases, the lowest values of S were obtained with $\mu = 0.01$, and that is the value we used for Figures 4.10, 4.11 and 4.12 —the shift in the equilibrium composition is more easily visualized in this case.

The different patterns observed in the three networks are a result of their size, the distribution of fitness values, and the similarities between environments, as well as the parameter μ. In order to properly discuss the results and their evolutionary consequences, we will first introduce a finite-population model, that will help us track population number as a function of p.

Our discrete-time, finite-population model is a modification of the classical Wright-Fisher model (Hartl et al., 1997). We start with a population of $N(0) = K$ individuals, distributed uniformly at random on the genotypes of the network. In each time step, every individual reproduces according to their fitness, so that

$$\begin{aligned} \Pr(O_i(t+1) = k_i \,|\, N_i(t)) &= \frac{(N_i(t) f_i(t))^{k_i}}{k_i!} e^{-N_i(t) f_i(t)}, \quad i = 1, \ldots, G. \\ N(t+1) &= N(t) + \sum_{j=1}^{G} O_i(t+1) \end{aligned} \tag{4.5}$$

where $O_i(t+1)$ is the offspring of node i at time $t+1$ and $N_i(t)$ is the number of individuals in the population at time t that have genotype i. In other words, reproduction follows a Poisson process in which every genotype reproduces independently. Note that the parental population does not die —the offspring is added to it. Every born individual will be a mutant with probability μ, and its genotype will be chosen uniformly at random among all the parent's neighbors. Then, death happens following a density-dependent law. Instead of choosing individuals to die, individuals are chosen to survive according to a binomial distribution with parameter $q = \exp(-N/K)$, which is 1 when $N = 0$ and decays quickly as the population grows, thus introducing a dependence on density modulated by parameter K. We assume K to be equal in both environments for simplicity. Finally, the environment changes with probability p, following the same dynamics discussed for our infinite-population model. In short, the model goes through four stochastic phases every time step: (a) birth, (b) mutation, (c) death (or survival) and (d) environmental change. We performed simulations of this model for different values of μ and K, varying p from 0 to

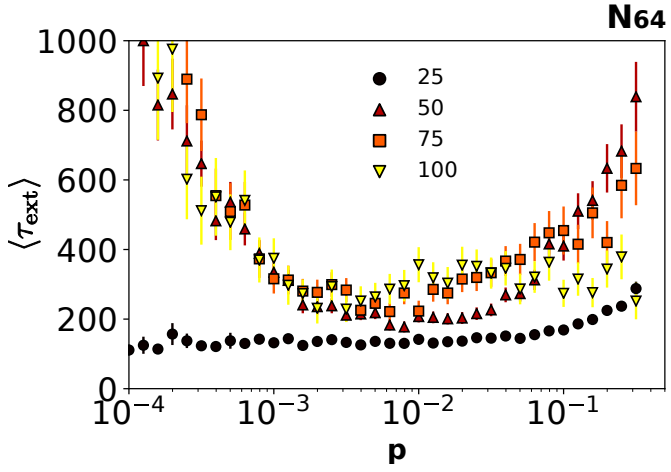


Figure 4.14: Effect of parameter K. Results for the simulations of our finite population models for N64, for $K = 25, 50, 75$ and 100. Note that the overall —noisy— dynamics remain the same, and that the value of K only affects the quantitative details. Markers represent sample means, and error bars represent the standard error of the mean. For these simulations, $\mu = 0.01$.

1. We studied the evolution of population size until it became extinct, and computed $\langle \tau_{\text{ext}} \rangle = \min\{t : N(t) \leq 0\}$, the average extinction time. For each combination of μ, K and p we performed 500 realizations of the process. The results are shown in Figure 4.13 (right panels). The effect of K, the parameter modulating death, is only quantitative and does not change the qualitative dynamics which we will discuss in the following (Figure 4.14).

As we can see in Figure 4.13, there is a value of p for which the average time to extinction is minimized. The simulations are noisy and quantitatively different in each network, but the overall dynamics is the same. Note that the y-axis is in logarithmic scale, and that changes in $\langle \tau_{\text{ext}} \rangle$ are sometimes subtler. This is a non-trivial result: we would expect that faster environmental changes would be best in order to eliminate a population, but that does not seem to be the case here. Our intuitive explanation for this result is the following: for very high values of p, and if it does not immediately become extinct, the population will remain concentrated on the overlapping regions —those regions of genotype space which are more or less fit in both environments (Figures 4.10, 4.11 and 4.12). As p decreases, the population has more time to expand in genotype space before a new environmental change occurs, and when this event finally occurs many individuals will suddenly become unfit: we expect the growth of the population to decrease significantly in that moment. Although the equilibrium composition of the population in the infinite case has not changed much, the population spends more time on those nodes with lower fitness, and therefore suffers more with each environmental change. As p decreases even more, the population spreads more and more in genotype space, and will suffer more strongly when the environment changes, but it will have more time to recover after this event, improving its chances of surviving the next environmental change. The magnitude of this population spread will depend on the evolution of S with p, as previously discussed, and this different evolution will no doubt be reflected in the differences observed in the three networks related to extinction times. For instance, the effect of p on $\langle \tau_{\text{ext}} \rangle$ is much less evident in the N32 network, and this may be caused by the high similarity between the two environments, as evidenced by the evolution of S (Figure 4.13 (lower left)).

Our results with finite populations suggest applications in the field of antibiotic resistance. Bacteria have developed resistance to all antibiotics

in clinical use (Payne et al., 2007), posing a serious threat to public health. As the development of new antibiotics has slowed down in recent years (Levy and Marshall, 2004), researchers look to new antibiotic administration strategies —what is usually termed antimicrobial stewardship (Fishman, 2006)— in order to solve this problem. Among these strategies, antibiotic cycling or sequential therapy is one of the most promising. The fundamental idea behind it is based on the fact that resistance evolution is due to exposition to the antibiotic. In the absence of the molecule, selective pressure disappears and the frequency of resistant strains decreases (Niederman, 2003). Recent experimental studies show that sequential therapy can be highly effective in controlling bacterial populations (Imamovic and Sommer, 2013; Kim et al., 2014; Fuentes-Hernandez et al., 2015; Roemhild et al., 2015).

Our multiplex framework contributes new theoretical support for sequential therapy, even in the absence of fitness cost associated with resistance —a plausible scenario (Rodríguez-Rojas et al., 2010). Here, one antibiotic treatment will represent one environment. In the presence of a given antibiotic, the population will expand through the neutral network due to mutations. When the environment changes, only those individuals belonging to the overlapping region of both neutral networks will survive. Depending on the mutation rate, there will be an optimal range of frequencies for environmental change which will maximize the eradication of the population (Figure 4.13). This prediction fits well with experiments (Peña-Miller et al., 2013). When exposed to combination therapies of two antibiotics, bacterial populations that were not completely eradicated recovered and grew very strongly, a fact that is consistent with our highest frequency scenario: if a population of bacteria is consistently being presented with two different antibiotics, only those bacteria that are resistant to both antibiotics will survive in the population, and they will grow unhindered. If, however, the population is exposed to an alternating set of antibiotics at lower frequencies of change the possibilities of exterminating the population will increase, as shown in Fuentes-Hernandez et al. (2015) and Roemhild et al. (2015) as well as in our results.

4.4 Summary

We have studied the prevalence of functional promiscuity in three different computational models of the genotype-phenotype map —RNA, Boolean GRNs and tOYLIFE. Our results show that functional promiscuity is very frequent in these models. RNA sequences are able to fold in many secondary structures. GRNs give rise to more than one expression pattern. tOYLIFE genotypes are able to metabolize several metabolites. The three definitions of promiscuity are different in each model, reflecting the differences in the corresponding definitions of phenotype. However, all of them show how easy it is for biologically-inspired models to generate secondary functions that can be easily exapted by evolution. Functional promiscuity is the norm, rather than the exception, in this kind of models. In tOYLIFE's case we should also note that it is remarkable that a genotype with only two genes is already able to metabolize so many different molecules. Although the rules in tOYLIFE are simple, and they do not necessarily correspond with real biology, the astounding complexity it shows hints at how easily cells in real life must perform promiscuous functions. Evolution at the cellular level must be a much more flexible process than we sometimes imagine. Additionally, we found that larger genotypes show, on average, less promiscuity, suggesting a constraining role of added genes in complex metabolism that we will need to explore further.

Our results also show that these three models are able to find many new phenotypes through promiscuity, and that a random walk through a neutral layer will discover new phenotypes in a linear fashion —in the case of RNA and Boolean GRNs— and almost discovering every possible phenotype —in tOYLIFE's case. These results point to the role of promiscuity as a facilitator of adaptation.

All three definitions of promiscuity focus on different aspects of cellular biology, and as such cannot be fully compared. All of them, however, suggest the expansion of the genotype network into the multiplex framework. Evolutionary dynamics in changing environments are more easily and intuitively studied in a multiplex genotype network: for metabolic systems, each layer would represent the presence or absence of a metabolite in the environment. For regulatory systems, there is a correspondence between layers and initial states. Finally, for molecular systems, we can think

of layers as molecular functions. In order to study the dynamical consequences of promiscuity, we have used two dynamical models to explore the effects of shifting environmental changes on the composition and survival of the population. We simulated both models in three networks, two of which were obtained from empirical fitness landscapes. Our results show that the equilibrium composition of the population changes slowly with the frequency of environmental change, and that there is a critical frequency at which the average extinction time of the population is minimized. This critical frequency depends on the mutation rate and the death rate, as well as the structure of the network and the correlation between fitness values in both environments. This result is coherent with experimental results in which bacterial populations are eliminated through changes of antibiotic treatment. We will need to explore these results further on in order to better understand the relationship between all these variables.

5

Spatio-temporal patterns in in toyLIFE

> "I love humans. Always seeing patterns in things that aren't there."
>
> *The Doctor*
> Doctor Who: the Movie (1996)

Cellular differentiation is a major invention of multicellular organisms. It allows them to divide the labor between different kinds of cells, thus optimizing resources and letting natural selection improve specific functions. But the cells in multicellular organisms all share the same genome, so this differentiation is achieved by regulatory means: some genes will be expressed in some cells, while others will not.

In developing organisms, the communication between cells and the enforcement of different regulatory programs in different cells invariably lead to spatio-temporal regulatory variation, that can be observed as patterns. In this chapter, we will see that toyLIFE has the potential to show regulatory patterns, thus allowing us to study their evolution.

5.1 Introduction

Gene regulatory networks (GRNs) act as integrators of signals from the environment (Alon, 2006). There are molecules that will trigger the expression of a given gene or genes: for instance, the presence of lactose in the environment will activate the genes related to its metabolism —the famous *lac* operon briefly mentioned in Section 4.3 (Ptashne and Gann, 2002; Alberts et al., 2014). Similarly, many other molecules trigger different responses in the cell. These signals are nothing else than a perturbation in the initial state of the regulatory dynamics. As we have seen before, different initial states in the GRN can be associated with different regulatory dynamics. Therefore, the change in the regulatory input will alter the dynamical attractor of the GRN.

In multicellular organisms, signal integration by GRNs is used in order to enable cellular differentiation. Because all cells in a multicellular organism share the same genome, the only way in which they can differ is in the way this genome is expressed. The modulation of gene expression is done through GRNs. The generation of different cellular lineages that will perform distinct functions inside the same organism is an instance of pattern formation (Phillips et al., 2012).

Different mechanisms for differentiation have been proposed, and not all of them fit into the previous discussion (Salazar-Ciudad et al., 2003; Phillips et al., 2012; Morelli et al., 2012). Some of these mechanisms do not need the influence of external regulatory signals, and make use of heterogeneities inside the cell to generate the pattern: differences in concentration in mRNAs or proteins within different regions of the cell may result in asymmetrical mitosis, which will lead to differential patterns of gene expression in the offspring. Other mechanisms are purely physical, in that the cellular response to a signal is not regulatory: for instance, cells can move inside the tissue, or change their form. In what follows, we will focus on those mechanisms involving the regulatory integration of external signals and cell-to-cell communication. Figure 5.1 shows some examples of the patterns obtained through some of these regulatory mechanisms: morphogen gradients, Turing-like patterns and lateral inhibition.

Morphogens are diffusible signals that influence gene expression in relation to their concentration. Cells are able to sense morphogens and

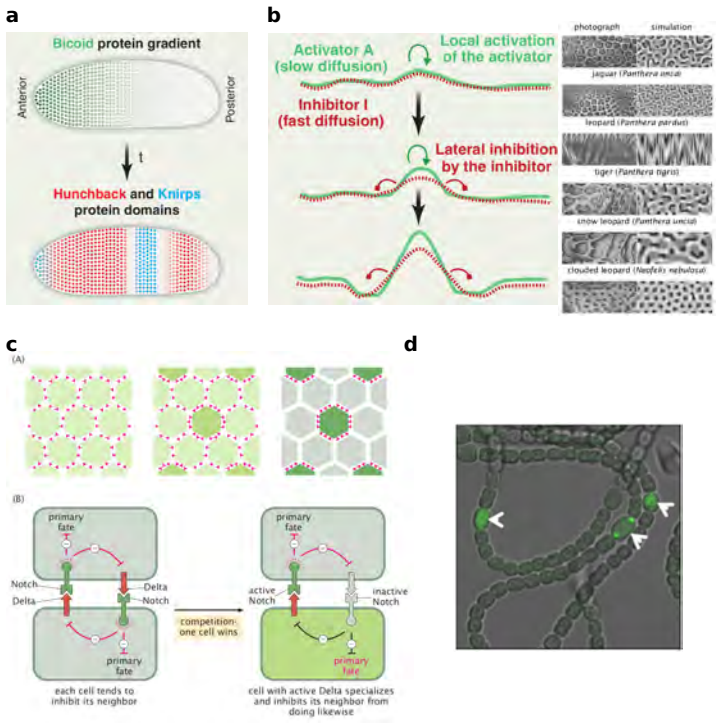


Figure 5.1: Mechanisms for pattern formation. **(a)** Morphogen gradients and development in *Drosophila*. The mRNA for the protein Bicoid is concentrated on the anterior end of the *Drosophila* embryo, leading to a difference in concentration along the embryo. This differential concentration triggers the expression of Hunchback and Knirps in different parts of the embryo. Figure from Morelli et al. (2012). **(b)** (left) Turing patterns are achieved when an activator and an inhibitor can diffuse through the tissue with different rates. Figure from Morelli et al. (2012). (right) Simulations with Turing's model can achieve very similar patterns to those observed in animal skin. Figure from Phillips et al. (2012). **(c)** Lateral inhibition in Notch-Delta can result in a fine-grained checkerboard pattern. The interplay between high and low concentrations of Notch and Delta achieves complex pattern without any need for diffusion (communication is cell-to-cell). Figure from Phillips et al. (2012). **(d)** Heterocyst (green cells) formation in *Anabaena* takes place every 10-12 cells. Here, pattern formation is coupled to the growth of the colony. Figure from Corrales-Guerrero et al. (2014). See Muñoz-García and Ares (2016) for more details.

develop different responses depending on their concentrations. Perhaps the most famous and simple example of morphogen is the Bicoid protein, whose mRNA is found at the anterior end of the *Drosophila melanogaster* embryo. When laying the egg, the mother leaves a high concentration of mRNA for the Bicoid protein, that will diffuse through the embryo, leading to an exponential decay in the concentration of Bicoid along the embryo (Figure 5.1**a**). The different levels of expression in Bicoid will activate and repress the expression of several genes in the *Drosophila*, starting the process of differentiation (Little et al., 2011).

Turing-like patterns arise when the patterns are the result of interactions between cells, without any external, spatially distributed signal —that is, all the cells in the tissue are in the same regulatory state at the beginning of the dynamics. Alan Turing was the first to propose a mechanism for the arising of these patterns (Turing, 1952). He hypothesized the existence of two different chemical compounds, an activator and an inhibitor, that could diffuse through the tissue and react with each other. The activator can activate both its own secretion and the inhibitor's. If we suppose that the inhibitor diffuses faster than the activator, then it will inhibit the production of the activator in the surrounding tissue, and we will end up with local peaks of the activator. Turing's mathematical analysis shows that this simple model is enough to generate a wide variety of patterns, from the tiger's stripes to the leopard's spots (Figure 5.1**b**). These patterns are called Turing-like because, although Turing's work is impressive and has fueled research over the decades, the mechanisms underlying most of these patterns is unlikely to correspond to Turing's proposed mechanism —for instance, the stripes and spots in animal skin are too large to be the result of diffusing molecules (Phillips et al., 2012). However, there is some experimental evidence that Turing's proposed mechanism is behind some observed patterns (Morelli et al., 2012).

Lateral inhibition is the basis of the Notch-Delta system, one of the most important mechanisms underlying animal development (Phillips et al., 2012). In this mechanism, one cell assumes a particular fate and then prevents neighboring cells from acquiring the same cell fate. The message is sent through Notch, a membrane-bound receptor that is activated by the Delta protein. The interplay between concentrations of Delta and Notch in different cells results in spatially inhomogeneous patterns, without any of

these proteins diffusing (Figure 5.1**c**). The interest of this mechanism lies in its potential to generate very fine-grained patterns, in which individual cells take fates that are different from their near neighbors. Lateral activation —in which one cell influences its neighbors to adopt the same fate as itself— is also a pattern-generating mechanism of relevance.

Finally, patterns can be coupled to tissue growth, such as in heterocyst-forming cyanobacteria (Figure 5.1**d**) and somite formation in vertebrates, leading to complex striped patterns (Morelli et al., 2012; Muñoz-García and Ares, 2016).

The prevalence and relevance of pattern formation mechanisms has fueled the interest to study their evolution. In particular, the study of the robustness and evolvability of cell patterns has been well studied by James Sharpe's group and his collaborators (Cotterell and Sharpe, 2010; Schaerli et al., 2014; Jiménez et al., 2015). For their studies, they model small GRNs with differential equations that track the concentration of protein products through time. Evolution in these models affect both the topology of the GRN and the parameters of the differential equations. Cotterell and Sharpe (2010) found that the same pattern could be obtained by multiple mechanisms —spatio-temporal dynamics associated to a group of genes. They observed that these mechanisms —that are obtained from several topologies— do not belong to a single neutral network in genotype space. In other words, in order to change the underlying mechanism that generates a given pattern, these genotypes have to go through other patterns. This insular disposition of clusters in genotype space is similar to what we observed in the metabolic phenotype in toyLIFE for small genotypes (see Chapter 3 in this thesis). Schaerli et al. (2014) classified these mechanisms and found the minimal topologies that were able to generate them. Jiménez et al. (2015) further showed that the evolvability of the pattern is dependent of the underlying mechanism that generates it.

In this chapter, we will show how a pattern-forming phenotype can be implemented in toyLIFE, and we will briefly study its robustness and evolvability. The interest of toyLIFE is two-fold: first, it allows us to directly study low-level evolution of protein sequences and promoter regions, in contrast to higher-level models of GRNs, that are restricted to studying the evolution of topologies. With toyLIFE we can study if some topologies are more easily generated than others, as well as their robustness and

evolvability. Second, toyLIFE automatically includes non-linear regulatory mechanisms, that are usually excluded from the models previously discussed.

Of course, these advantages come at a cost. toyLIFE is not a model for real biology, and therefore the conclusions we can extract from it must be treated with caution. However, we are confident that the insights gained from *playing* with toyLIFE can give rise to new ideas on molecular evolution.

toyLIFE is also limited because it is a Boolean model of regulation. When the pattern-formation mechanism is inherently continuous, Boolean models do not capture them easily. For instance, a morphogen gradient such as the one shown in Figure 5.1**a** is not easy to implement with toyLIFE —although there are tricks around the constraints imposed by the Boolean formalism.

The results shown in this chapter are preliminary, and are just intended to show the potentialities of toyLIFE in studying spatio-temporal regulatory patterns in multicellular organisms. Instead of performing the same analyses we have done for the metabolic genotype-phenotype map in Chapter 3, we will just do a shallow exploration of evolvability and robustness.

5.2 Definition of phenotype

We will only use the regulatory function of toyLIFE, ignoring the existence of metabolism. In order to study patterns in a multicellular organism, we studied a row of 31 cells in a row (see Figure 5.2). The number of cells is arbitrary. We could —and have— studied other row sizes, but we will restrict our discussion to our results in a 31-cell row: they are large enough to capture all complex patterns that two-gene toyLIFE genotypes are able to generate, and their small size allows for clear visualization.

We also restricted ourselves to genotypes with 2 genes. Again, we could study the phenotype with larger genotypes, but we will not do so in this chapter.

For each cell, the instructions described in Chapter 2 define the regulatory output given any input. However, we need to specify the connections between cells. For simplicity, we will assume that some of the expression products —toyProteins— will be able to diffuse to the adjoining cells

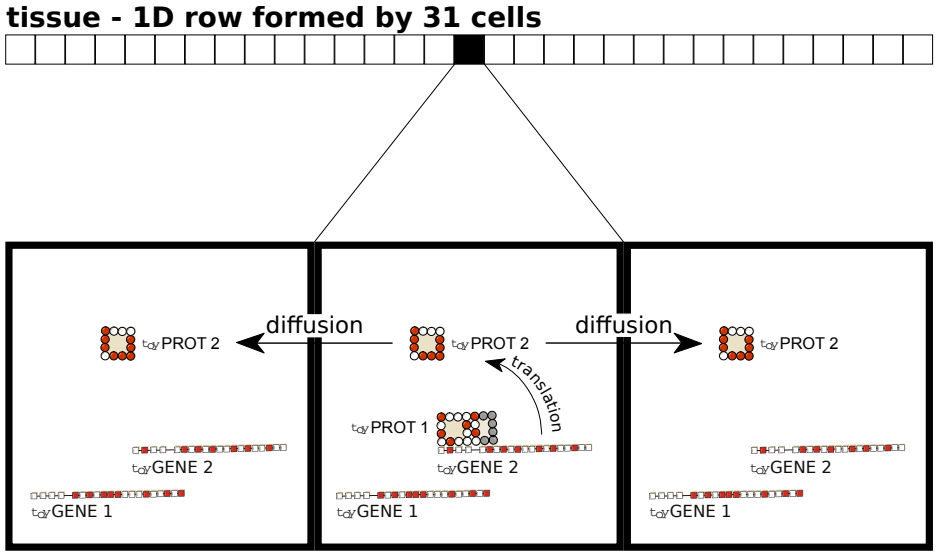


Figure 5.2: Pattern formation phenotype in toyLIFE. We consider a one-dimensional row formed by 31 cells. The figure illustrates one example of how this multicellular phenotype works. Suppose that toyProtein 1 is present in the central cell of the row after a certain time step. Then toyProtein 2 is activated, and it will diffuse to the neighboring cells, affecting their output.

—this will lead to lateral inhibition and lateral activation scenarios. As a result, the input set of cell c_i in time $t+1$ will be affected by the output states of cells c_{i-1} and c_{i+1} in time t —as well as its own (Figure 5.2). We will further assume that there is enough toyProtein to stay inside the cell and diffuse to the adjoining ones. For the cells at the beginning and end of the row, we can impose different conditions. The one we have chosen here is a non-periodic one: cell c_0 will be affected by itself and cell c_1, and cell c_S will be affected by itself and c_{S-1} —remember that $S = 30$ in our case.

We have studied four different communication scenarios: (1) only the first toyProtein is able to diffuse; (2) only the second toyProtein is able to diffuse; (3) both toyProteins are able to diffuse, but the toyDimer is not; and (4) both the two toyProteins and the toyDimer (when formed) are able to diffuse.

In summary, our toyLIFE phenotype will be discrete in time and space, with a discrete number of states —four: (0) no toyProtein is expressed,

(1) toyProtein 1 is expressed, (2) toyProtein 2 is expressed and (3) both toyProteins are expressed, forming a toyDimer or not —for our purposes it is irrelevant. In this sense, our phenotype is a cellular automaton with 4 states (see Wolfram (2002) for a thorough exploration of cellular automata). Cellular automata are easily described by the output they produce given an input. Because the input of a cell is formed by itself and its adjoining cells, and because each of them can be in 4 states, the number of input states is $4^3 = 64$. The number of possible cellular automata is, therefore, $4^{64} \sim 3.4 \times 10^{38}$. Remember that the number of $g = 2$ genotypes is 5.5×10^{11}, a tiny number in comparison with the former, so it is obvious that toyLIFE genotypes with two toyGenes will be able to generate only a minute fraction of the cellular automata.

The cellular automata, in turn, will be uniquely determined by the logic function of a genotype, once we take into account the two additional input states: toyProtein 1 plus toyDimer, and toyProtein 2 plus toyDimer — which were not considered in previous chapters (Figure 5.3**a**). There are $1,191$ different logic functions, not all of them equally abundant (Figure 5.3**b**). Considering all four communication scenarios, we obtain $4,764$ different logic functions. In total, they generate $1,535$ cellular automata. Again, we see a huge degeneracy in the genotype-phenotype map thus defined, and an enormous skewness in the distribution of phenotype sizes (Figure 5.3**c**).

5.3 Diversity of patterns

In order to show the diversity of patterns generated by these cellular automata, we computed the output obtained from a simple input: the one-time activation of toyProtein 1 in the middle cell of the row. We then studied the evolution of the formed pattern for 100 time steps. Some automata generated the same pattern, and in the end we obtained 494 spatio-temporal patterns in this setting —this number contains all patterns observed for larger rows or longer times, however. Figures 5.4 to 5.10 show some selected patterns that illustrate the diversity and complexity of the toyLIFE regulatory phenotype. The horizontal axis represents the spatial dimension: each column represents one cell in the row, from $c = 0$ (left) to $c = 30$ (right).

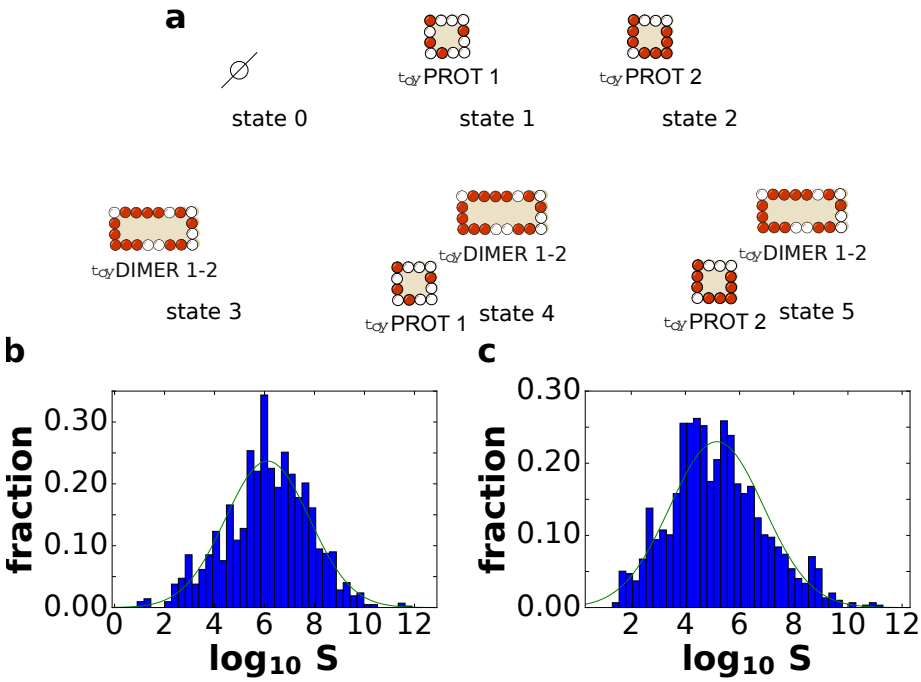


Figure 5.3: Cellular automata are uniquely determined by the logic function. (a) Because toyProteins can diffuse between cells, the logic function has two new input states that were not discussed in previous chapters. **(b)** Histogram of abundances of logic functions. The distribution is somewhat close to a log-normal. **(c)** Histogram of abundances of cellular automata. As in **(b)**, the distribution is reminiscent of a log-normal.

The vertical axis is the temporal dimension, and each row represents one time step, from $t = 0$ (top) to $t = 99$ (bottom).

The most common phenotype (not shown) is a homogeneous one: although there is an initial signal, the whole row ends up in the same state very quickly. Figure 5.4 shows some repeating patterns. Pattern 114 —the numbers are arbitrary, and just reflect the id assigned to them by the computer program— is just a stable response to a one-time signal: toyProtein 1 is activated and remains expressed forever, while the neighboring cells express toyProtein 2 independently, without any interference. Pattern 8 is similar, but the generated stripe is wider: it is formed by three cells expressing toyProtein 2. The remaining cells do not express anything. We

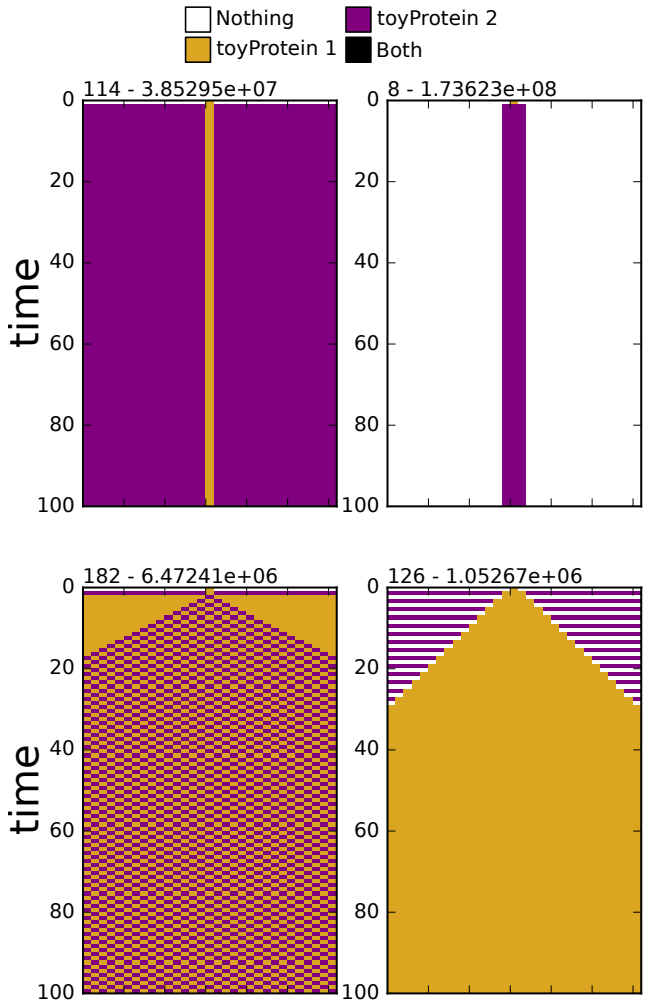


Figure 5.4: Patterns in toyLIFE (1). Four patterns obtained from two-gene toy-LIFE genotypes, after activating the expression of toyProtein 1 in the central cell. The horizontal axis represents the spatial dimension, from cell $c = 0$ (left) to cell $c = 30$, while the vertical axis represents the vertical axis, from $t = 0$ (top) to $t = 99$ (bottom). The first number on top of the pattern is the id assigned to the pattern by the computer program, while the second number is the number of genotypes that generate said pattern. The color code represents what is being expressed in each cell in each time step (see legend). See text for more details.

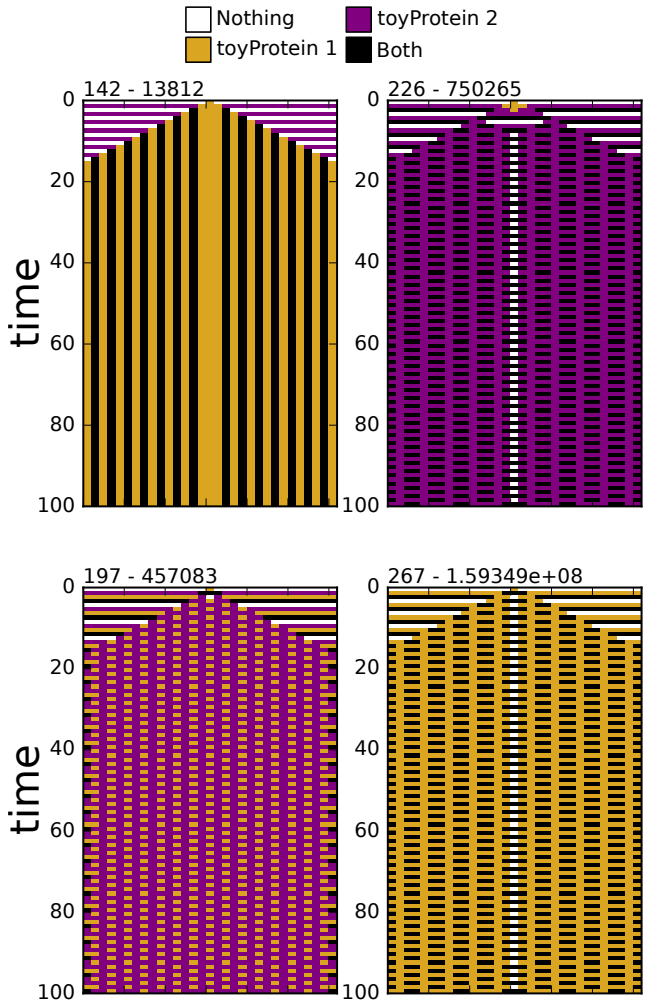


Figure 5.5: Patterns in toyLIFE (2). Four patterns obtained from two-gene toy-LIFE genotypes, after activating the expression of toyProtein 1 in the central cell. The horizontal axis represents the spatial dimension, from cell $c = 0$ (left) to cell $c = 30$, while the vertical axis represents the vertical axis, from $t = 0$ (top) to $t = 99$ (bottom). The first number on top of the pattern is the id assigned to the pattern by the computer program, while the second number is the number of genotypes that generate said pattern. The color code represents what is being expressed in each cell in each time step (see legend). See text for more details.

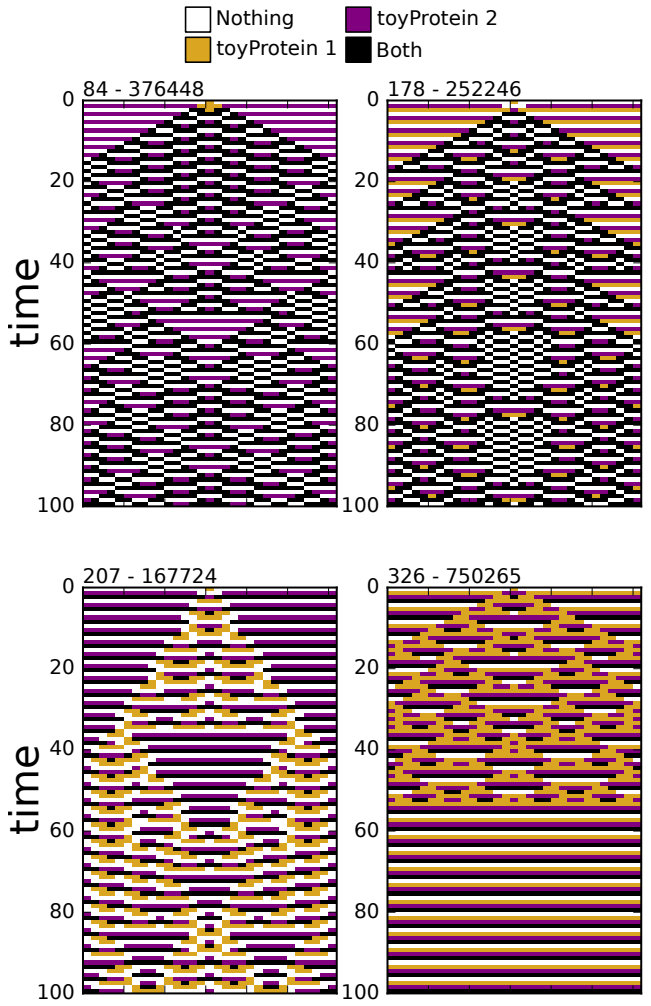


Figure 5.6: Patterns in toyLIFE (3). Four patterns obtained from two-gene toy-LIFE genotypes, after activating the expression of toyProtein 1 in the central cell. The horizontal axis represents the spatial dimension, from cell $c = 0$ (left) to cell $c = 30$, while the vertical axis represents the vertical axis, from $t = 0$ (top) to $t = 99$ (bottom). The first number on top of the pattern is the id assigned to the pattern by the computer program, while the second number is the number of genotypes that generate said pattern. The color code represents what is being expressed in each cell in each time step (see legend). See text for more details.

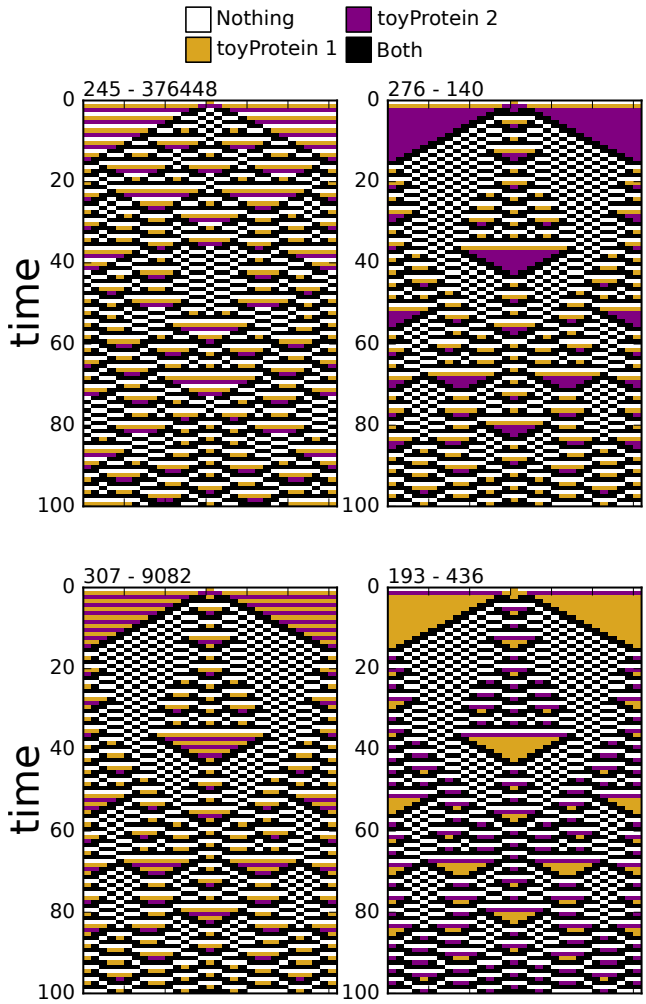


Figure 5.7: Patterns in toyLIFE (4). Four patterns obtained from two-gene toy-LIFE genotypes, after activating the expression of toyProtein 1 in the central cell. The horizontal axis represents the spatial dimension, from cell $c = 0$ (left) to cell $c = 30$, while the vertical axis represents the vertical axis, from $t = 0$ (top) to $t = 99$ (bottom). The first number on top of the pattern is the id assigned to the pattern by the computer program, while the second number is the number of genotypes that generate said pattern. The color code represents what is being expressed in each cell in each time step (see legend). See text for more details.

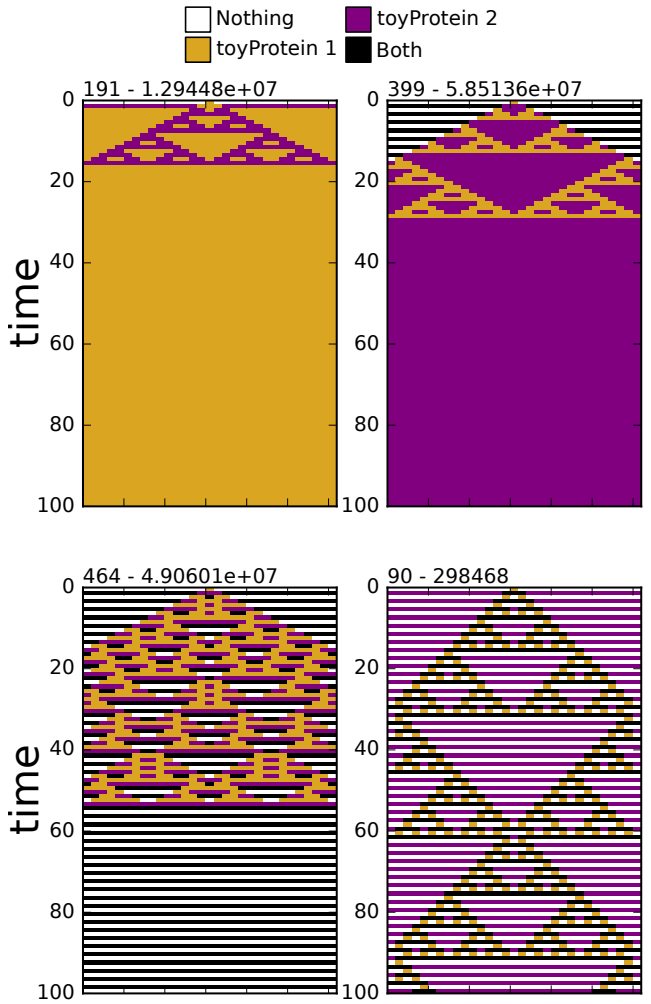


Figure 5.8: Patterns in toyLIFE (5). Four patterns obtained from two-gene toy-LIFE genotypes, after activating the expression of toyProtein 1 in the central cell. The horizontal axis represents the spatial dimension, from cell $c = 0$ (left) to cell $c = 30$, while the vertical axis represents the vertical axis, from $t = 0$ (top) to $t = 99$ (bottom). The first number on top of the pattern is the id assigned to the pattern by the computer program, while the second number is the number of genotypes that generate said pattern. The color code represents what is being expressed in each cell in each time step (see legend). See text for more details.

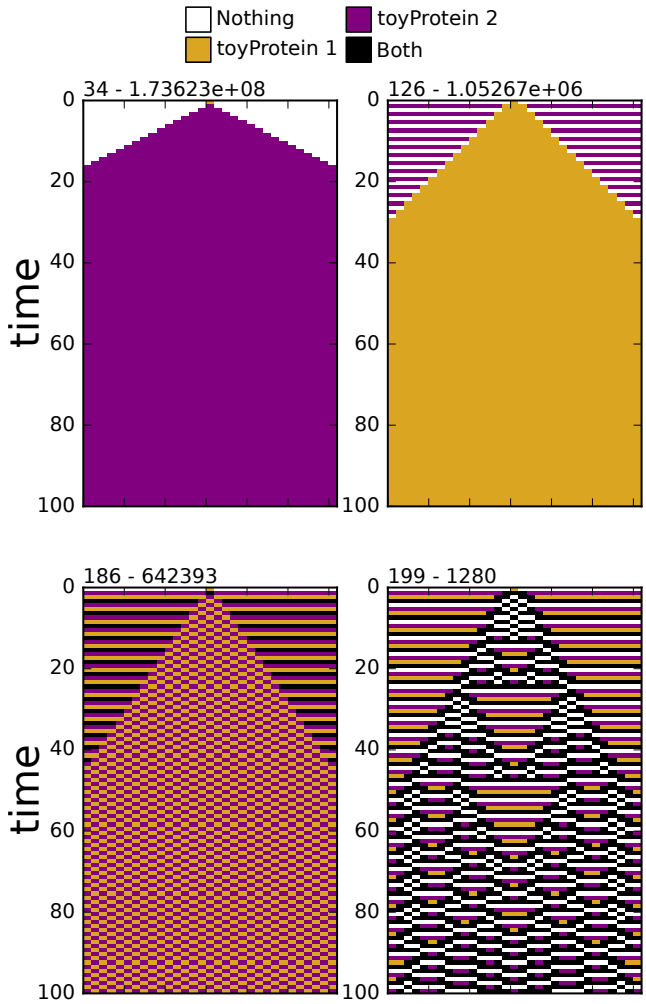


Figure 5.9: Patterns in toyLIFE (6). Four patterns obtained from two-gene toy-LIFE genotypes, after activating the expression of toyProtein 1 in the central cell. The horizontal axis represents the spatial dimension, from cell $c = 0$ (left) to cell $c = 30$, while the vertical axis represents the vertical axis, from $t = 0$ (top) to $t = 99$ (bottom). The first number on top of the pattern is the id assigned to the pattern by the computer program, while the second number is the number of genotypes that generate said pattern. The color code represents what is being expressed in each cell in each time step (see legend). See text for more details.

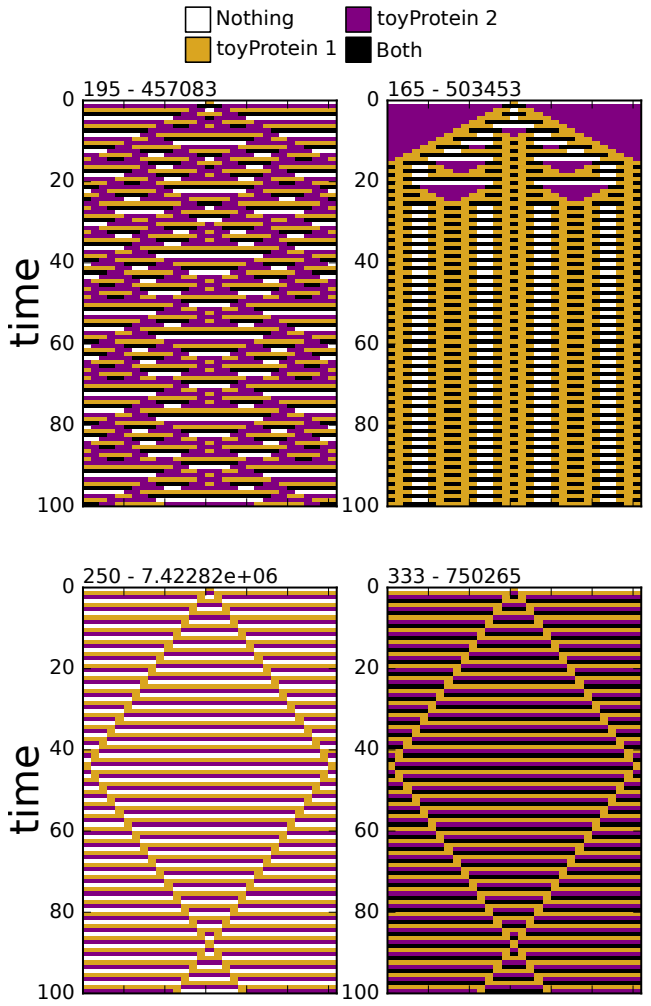


Figure 5.10: Patterns in toyLIFE (7). Four patterns obtained from two-gene toyLIFE genotypes, after activating the expression of toyProtein 1 in the central cell. The horizontal axis represents the spatial dimension, from cell $c = 0$ (left) to cell $c = 30$, while the vertical axis represents the vertical axis, from $t = 0$ (top) to $t = 99$ (bottom). The first number on top of the pattern is the id assigned to the pattern by the computer program, while the second number is the number of genotypes that generate said pattern. The color code represents what is being expressed in each cell in each time step (see legend). See text for more details.

can easily see that the automata that generate pattern 114 and pattern 8 will be able to generate stable striping patterns in response to spatially distributed signals from the environment. Note, moreover, that these patterns are not particularly rare: pattern 114 is generated by 3.85×10^7 genotypes with two genes, and pattern 8 is generated by 1.74×10^8 genotypes. The remaining two patterns in Figure 5.4 represent advancing patterns, the result of lateral inhibition and activation. In pattern 182, the starting signal is carried to the rest of the row, and all the cells end up alternating the expression of toyProtein 1 and toyProtein 2, forming a checkerboard pattern. In pattern 126 the signal is amplified and carried to the rest of the row, and the whole tissue ends up expressing toyProtein 1. Both patterns are generated by more than one million genotypes.

Figure 5.5 shows patterns with stable stripes. In all of them, the initial signal is transported to the rest of the tissue, and the cells end up expressing a slightly different alternating pattern. Pattern 142 is somewhat rare —only $13,812$ genotypes express it— but it is the best example of a standard alternating pattern. Except from a central three-cell wide stripe, the rest of the tissue alternates between expressing toyProtein 1 and expressing both toyProteins. Pattern 226 also presents stable stripes, in this case expressing toyProtein 2, with spaces in which the neighboring cells alternate between expressing toyProtein 2 and expressing both toyProteins. Note that the central cell alternates between not expressing anything and expressing both toyProteins. Pattern 197 also generates a stable alternating pattern with toyProtein 2, and the neighboring cells alternate between expressing toyProtein 1 and 2. Finally, pattern 267 is similar to pattern 226, but with different toyProteins involved. All of these patterns represent examples of lateral inhibition and activation.

Figure 5.6 shows stranger patterns, with a long temporal period. In all of them, the initial signal is transported to the rest of the row, but the interaction between neighboring cells leads to interference and unstable patterns. In pattern 84, a complex non-repeating pattern eventually achieves equilibrium: a stable pattern that repeats each 22 temporal steps. Equilibrium is reached because of the border effect from the finite length of the row. For larger rows, this pattern would take longer to collapse, and the period would be larger. For example, if we repeat the simulation with a row of 99 cells, the period of the same pattern would be 116. Moreover,

if the row was infinitely long, no period would be observed. In pattern 178, equilibrium is not reached after 100 time steps, but the final period of the equilibrium pattern is 30. For pattern 207, the equilibrium pattern is nowhere to be seen in the first 100 time steps, while in pattern 326 it is reached after roughly 50 time —it is a spatially homogeneous pattern that alternates temporally every 4 time steps, going through every state possible.

These complex, long-periodic patterns are probably not very useful for multicellular organisms, but they are shown here as proof of the potential of toyLIFE genotypes.

Figure 5.7 shows four very similar patterns, more or less rare, that show fractal structures. The upside-down triangles in patterns 276, 307 and 193 are repeated over and over again, albeit in different sizes and positions. Pattern 245 also shows repeating structures, that can be thought of as multicolored triangles. No temporal period is discerned in either of them. When they finally get to equilibrium, their periods are 62 for pattern 245 and, surprisingly, 2 for patterns 276 and 307 and 3 for pattern 193. This shows that these cellular automata can take very long times to reach equilibrium, even though they are not very complex.

Figure 5.8 shows more fractal examples, in this case four versions reminiscent of Sierpinski's triangle, one of the most well-known fractal structures. Note that patterns 191, 399 and 464 all reach homogeneous spatio-temporal patterns, in which all cells have the same state permanently. Interestingly, these three patterns are not uncommon: each are mapped by more than 10^7 genotypes. Pattern 90 is an instance of a repeating triangle, with temporal period 60.

Figure 5.9 shows that the response to the signal can be transmitted through neighboring cells at different speeds. Because we only allow for diffusion to adjacent neighbors, the maximum speed of the message is one cell per time step, which is what pattern 34 accomplishes. Because the row has 31 cells and the signal starts in the middle, the signal traverses 15 cells in each direction, in 15 time steps. In pattern 126, the signal advances one cell each two time steps, so the speed is $1/2$. Although the front in pattern 186 is more complicated, if we focus on the cells expressing toyProtein 2, we can see that the pattern advances one cell every three time steps, and thus the speed is $1/3$. Finally, in pattern 199 the speed is $1/4$: focusing on

the cells expressing both toyProteins, we see that the pattern advances one cell each four time steps.

Finally, Figure 5.10 shows four additional patterns that add to the complexity already shown. Pattern 195 is a strange pattern that repeats itself every 20 time steps, creating fascinating structures in the process. Pattern 165 is another example of a striped pattern, with more complex stripes. Patterns 250 and 333 are examples of an advancing front that rebounds against the edge of the row. This pattern gives information on the length of the row, as it will take longer to come back the longer the row is. Measuring structures' length is not a trivial issue —cells do not have eyes!— so this kind of patterns could potentially be very useful.

It is remarkable that two-gene genotypes are able to generate such wide diversity of patterns, from stable stripes to fractal structures to rebounding fronts. It is also remarkable that these patterns can be generated with toyLIFE without any additional modification. Note also that most of these patterns are not rare at all, and that the potential of toyLIFE genotypes to generate them seems unrestricted. Real genomes are much more complex than toyLIFE, so it is no wonder that we observe so many wonderful and complex patterns in nature: it is just easy to generate them.

5.4 Robustness and evolvability

We will restrict ourselves to a very shallow exploration of robustness and evolvability of the pattern-forming genotype-phenotype map in toyLIFE. For robustness, we sampled 10^7 genotypes that did not generate the trivial logic function —that is, something had to be expressed at some point. We then studied all their 40 mutants and computed their phenotypes as described above —that is, two genotypes have the same phenotype if they generate the same pattern after the initial signal in the central cell. Robustness is the same as defined in Chapter 3. The results are shown in Figure 5.11**a**. Note that robustness in this genotype-phenotype map is much higher than in the metabolic map studied in Chapter 3. This is in part explained by the smaller number of phenotypes, and in part because the regulatory function is much more robust than the metabolic one. Also, many of these genotypes are non-viable when considered under the metabolic genotype-phenotype map, so the lower robustness observed in Chapter 3

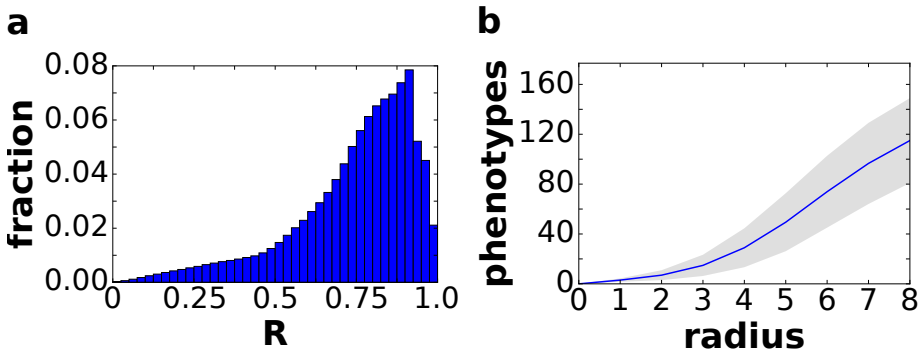


Figure 5.11: Robustness and evolvability. (a) Histogram of robustness for 10^7 randomly sampled genotypes. **(b)** Shape space covering. We randomly sampled 100 genotypes. For all of them, we computed the phenotypes for all genotypes in a radius of distance 8 around the given genotype. The figure shows the average (blue line) plus minus one standard deviation (gray area).

does not translate well to these genotypes, as the set we are considering is different.

As for evolvability, we studied shape space covering: randomly sampling one genotype, we studied its neighborhood up to a radius of 8 mutations, and counted how many patterns were found inside that ball. We repeated that process for 100 genotypes. The results are shown in Figure 5.11**b**. Note that these genotypes are not as evolvable as the metabolic ones, but either way up to 25% of the phenotypes is discovered in less than 9 mutations away from a randomly sampled viable genotype.

5.5 Summary

This brief exploration of the pattern-forming phenotype in toyLIFE shows that this simple model already shows the potential to generate many complex and intricate patterns, that are both highly robust and evolvable.

In this chapter we have restricted ourselves to a very specific pattern-formation scenario, namely, that of a single signal expanding through a tissue via lateral inhibition or activation. However, different patterning mechanisms could be explored in toyLIFE. Morphogen gradients, although difficult to implement due to the binary nature of toyLIFE, could be imple-

mented through a temporal signal: higher concentrations of the morphogen would be translated as more frequent inputs to the cells, while lower concentrations would mean less frequent inputs. Reaction-diffusion could be implemented by adding a lag time for the diffusion and disappearance of the toyProteins in the cells. Finally, we could also experiment with growing rows —and 2-D tissues!

6

Adaptive multiscapes: An up-to-date metaphor to visualize molecular adaptation

"I've seen things you people wouldn't believe"

Roy Batty, replicant
Blade Runner (1982)

The fitness landscape is arguably the most enduring and successful metaphor of evolutionary theory. The image it conjures, of populations climbing to higher fitness peaks, is one of the most straightforward ways to translate Darwin's insight into natural selection. However, as mentioned in Chapter 1, we have learned many things since Darwin's time, and the fitness landscape metaphor is not able to include all of them. Motivated by the results obtained throughout this thesis, and in combination with recent literature, in this chapter we present a new framework, adaptive multiscapes, that tries to update the fitness landscape metaphor for our modern times.

6.1 Introduction

The fitness landscape metaphor was conceived by Sewall Wright (1931) as a way to present his work in a non-mathematical way at the Sixth International Congress of Genetics. Wright envisioned evolution as a movement —in genotype space— from one fitness maximum, or *peak*, to another, traversing fitness *valleys* with the help of genetic drift (that is, roughly, the outline of his shifting balance theory). That his two-dimensional representation of genotype space was an over-simplification did not escape Wright, who was worried that the number of fitness maxima was too large or that the static view did not capture the effect of environmental changes (Wright, 1932). Nonetheless, the idea of populations moving on a physical landscape following "natural" directions of movement was extremely suggestive, and shaped evolutionary thinking for the next century (Pigliucci, 2008; Svensson and Calsbeek, 2012).

We can catch a glimpse of how influential the image of populations moving uphill in the fitness landscape is by taking a look at recent publications studying empirical fitness landscapes (Poelwijk et al., 2007; De Visser and Krug, 2014; De Vos et al., 2015; Steinberg and Ostermeier, 2016): often only monotonically increasing fitness paths are explored as evolutionarily relevant, and fitness peaks are studied on the basis of their accessibility through this kind of paths. Advances in our knowledge of molecular genotype-phenotype maps —to which this thesis pretends to be a humble contributor— suggest that there is more to evolution than this simplified picture, adding to Wright's original worries.

The first important topographical element missing from the naïve picture of fitness landscapes are ridges, or surfaces of equal fitness. In a high-dimensional genotype space, ridges correspond to neutral networks, which have been reviewed extensively throughout this thesis. Connected ridges appear if, on average, genotypes yielding the same phenotype have more than one neutral neighbor (Gavrilets and Gravner, 1997). Not only computational models of the genotype-phenotype map, such as those discussed in this thesis, but also many empirical studies (Eyre-Walker and Keightley, 2007; Schultes and Bartel, 2000; Bloom et al., 2007; Koelle et al., 2006) have shown the ubiquitousness of neutral networks, fulfilling Maynard Smith's condition for the navigability of genotype spaces (Maynard

Smith, 1970). A first attempt to include the existence of neutral networks —or ridges— in the existing framework was made by Sergey Gavrilets (1997), with his *holey adaptive landscapes*. However, these landscapes still suffer from the misleading representation of genotype space, in which genotypes appear closer to each other than they really are.

Additionally, the uneven size of phenotypes is also unrepresented in fitness landscapes. As we have seen throughout this thesis, most phenotypes are very rare, with most genotypes mapping onto a small set of phenotypes —a property observed in all models of the genotype-phenotype map (see Section 1.2). The mutual accessibility of phenotypes is, as a consequence, highly asymmetric: it is much easier to access a frequent phenotype from a rare one than the other way around. This asymmetry has been shown to affect the outcome of evolutionary dynamics, preventing the fixation of a fitter phenotype if it is rare enough (Schaper and Louis, 2014). Large phenotypes are, additionally, much more robust to mutations (Aguirre et al., 2011; Greenbury and Ahnert, 2015) and are easily accessed from each other by point mutations (Schultes and Bartel, 2000; Bloom et al., 2007; Koelle et al., 2006).

Finally, as we saw in Chapter 4, genotypes tend to be functionally promiscuous, that is, they express different phenotypes under different circumstances —the genotype-phenotype map is a many-to-many correspondence. In other words, phenotype —and, therefore, fitness— is a function of the environment. In Chapter 4, we saw promiscuity in RNA, GRNs and toyLIFE, and discussed some literature that had documented it for proteins (Aharoni et al., 2005; Piatigorsky, 2007) or metabolic models (Barve and Wagner, 2013). The static picture of fitness landscapes, associating one value of fitness to each genotype, misses this phenomenon. Moreover, there is a link between neutral networks and accessibility to new functions through functional promiscuity. For instance, experiments have shown that we can improve a protein's secondary function by point mutations that do not alter the primary function (Aharoni et al., 2005), allowing an heterogeneous population to adapt to a new environment in a faster way.

All of these features can —and indeed, need to— be combined into a new image of the evolutionary process, one that helps us to think about adaptive dynamics, but also to better communicate evolutionary ideas to non-specialists: the same need that prompted Wright to develop the fit-

ness landscape metaphor. We propose a renovated picture, that we term *adaptive multiscapes*. It contains some of the overall traits of Wright's metaphor, including details from subsequent re-formulations, but also incorporates the presence of neutral networks, the asymmetry in phenotype sizes and accessibilities, the absence of visual distance between genotypes and functional promiscuity.

6.2 Adaptive multiscapes

We are looking for a visual metaphor that integrates the three elements that we have discussed so far with the features of Wright's landscape. Thus, we need to include information on neutral networks and their uneven size distribution, the asymmetric accessibility between phenotypes, functional promiscuity, and the relationship between fitness and adaptation.

Figure 6.1 summarizes the main elements of adaptive multiscapes. First, we represent genotype space as a network (Figure 6.1**a**): genotypes are represented by nodes, and two nodes are connected if there is a mutation that transforms the first genotype into the second. In the literature, and throughout this thesis as well, we have only considered point mutations as valid mutational moves, but that is not a requirement of this representation: we could easily include duplications, deletions, or even horizontal gene transfer in the picture. The links between nodes introduce a natural definition of distance, which is unrelated to visual distance on the picture: two genotypes are not closer to each other if they are closer in the picture, only through the connections that join them in the network. Second, genotype space can be partitioned, as we saw in previous chapters, into disjoint neutral networks, each related to a phenotype (Figure 6.1**b**). Given an environment, each genotype is mapped to a single phenotype. Last, in Figure 6.1**c** we join these elements into one picture. Phenotypes are now represented by one node, whose radius is proportional to its size. The inner complexity of each phenotype is left out of this picture. Emphasis is put on the connections between phenotypes: the network of phenotypes is a directed one, and there is strong asymmetry between the links: going from a small phenotype to a large one is easier than the reverse move. But note that almost every phenotype is connected to each other. Fitness is included as a color code.

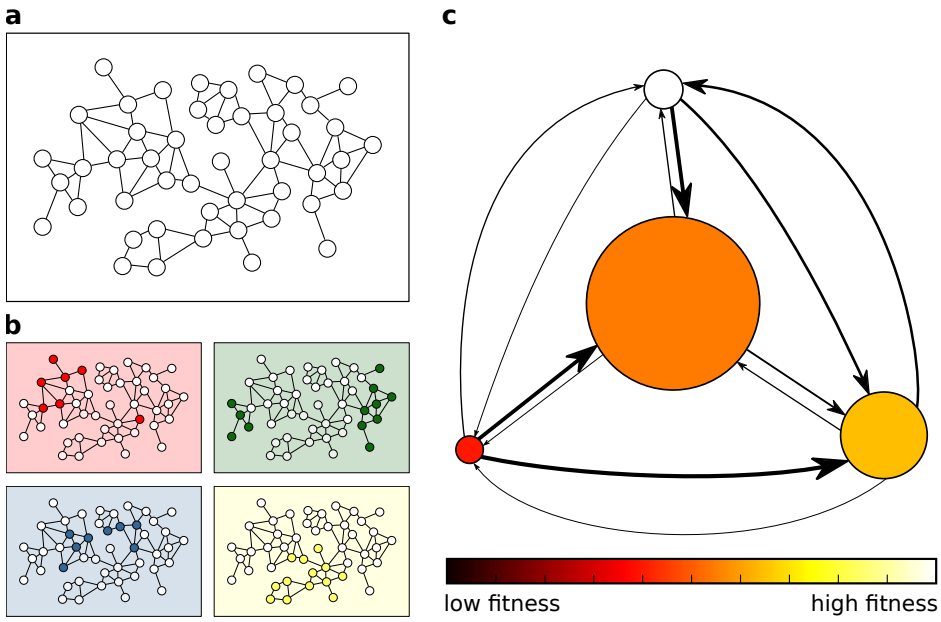


Figure 6.1: Basic elements in the construction of the adaptive multiscape metaphor. (a) Schematic representation of genotype space as a network of genotypes. Two genotypes are connected if they are one mutation away from each other. **(b)** In a given environment, genotype space can be partitioned into different networks corresponding to different phenotypes. In these new networks, we only consider genotypes expressing the same phenotype (in the figure, having the same color). **(c)** Synthetic representation of neutral networks, or phenotypes, as circles with radii proportional to their size —the number of genotypes that map into that particular phenotype. Two phenotypes are connected if we can mutate from the first into the second. The thickness of the arrows represents the likelihood of reaching the target phenotype from the initial phenotype. These links are not symmetrical (see text). Each phenotype has an assigned fitness in a given environment, represented here through a color code.

This last picture is highly dependent on the environment. Different phenotypes will have different sizes in different environments, because genotypes that express one phenotype in the first environment may express a different one in the second. Also, phenotype fitness can change dramatically: fitter phenotypes in one environment can be disadvantageous in another. Thus, we need to include the different environments as different

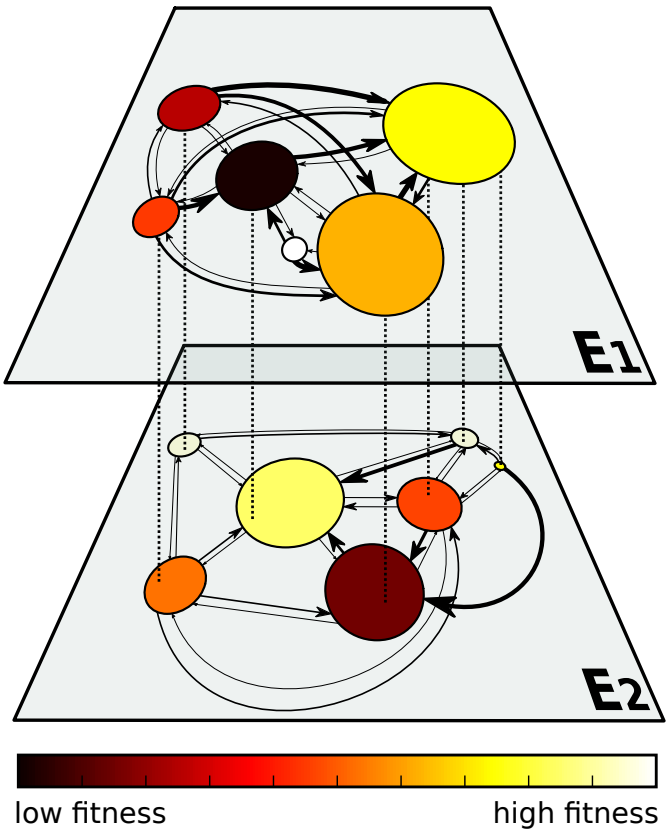


Figure 6.2: Genotypes can express different phenotypes in different environments. We represent two different environments (**E1** and **E2**) as two layers of a multiplex network. Dashed lines represent instances of functional promiscuity, where genotypes are viable in each environment, perhaps expressing different phenotypes. These genotypes belong to different neutral networks in those environments, with different fitness values. For clarity, only a subset of all possible connections between phenotypes is represented.

layers in a multiplex network (see Section 4.1.2). This new picture, shown in Figure 6.2, is the visual metaphor of adaptive multiscapes.

6.3 Population dynamics on adaptive multiscapes

Adaptive multiscapes help us intuitively understand evolutionary dynamics. We will now discuss how this intuition is built, before exploring some concrete examples.

Natural populations are finite, and therefore they cannot explore all of genotype space at once. Typically, one population will be exploring but a fraction of any phenotype. This limit is not a hindrance for evolution, as the population can easily move inside a phenotype through neutral mutations, and because any two common phenotypes are connected —through the shape space covering property already discussed in Chapters 1 and 3. The size of a population and its mutation rate have a direct effect on the time spent in a given phenotype, and on the probability to find new functions. Also, the size of the phenotype will affect the structure of the population: large phenotypes are more robust (again, see Chapters 1 and 3), and therefore populations will be more diverse, and more evolvable (Wagner, 2011).

Populations can be —qualitatively— represented in adaptive multiscapes as subsets of a phenotype, moving inside it and among different phenotypes. If the population is larger, it will "fill" more of the phenotype. As population size decreases, neutral drift becomes more important (Hartl et al., 1997) and trajectories inside a phenotype become less deterministic. When populations first access a phenotype, they do so through a small number of genotypes, and therefore tend to be highly homogeneous. After some time, they explore the neutral network and population diversity grows (Huynen et al., 1996). Eventually, if no fitter phenotype is found in the process and the environment does not change, the population stabilizes in the regions of maximal robustness inside the neutral network.

We have highlighted the relevance of neutrality in our discussion of dynamics so far. It is perhaps relevant to note that neutrality is absent from the early fitness landscape picture. Adaptive multiscapes naturally include this property, that promotes evolvability and diversity in populations. Because of the intricate dynamics inside a neutral network, a population can spend long times inside a phenotype (Manrubia and Cuesta, 2015): populations can get more and more "trapped" inside a phenotype the longer they stay in it, making adaptive moves more and more unlikely. Our adaptive multi-

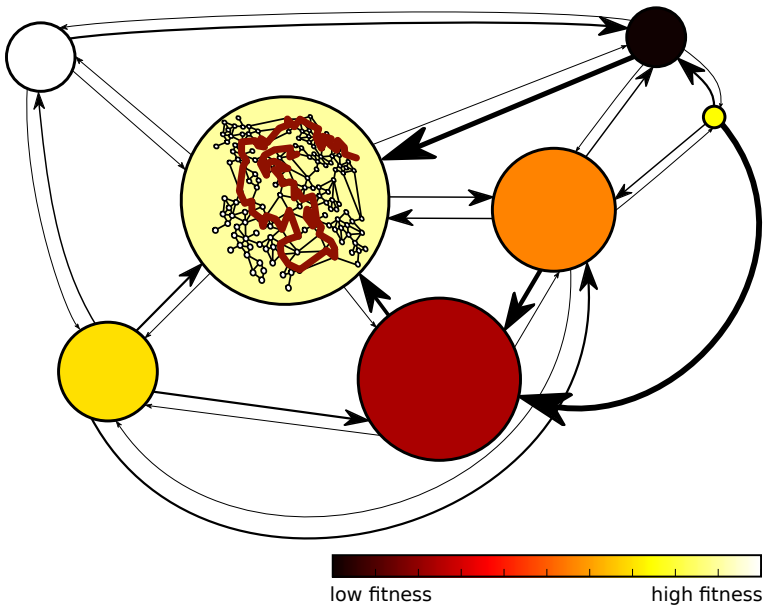


Figure 6.3: Population dynamics on adaptive multiscapes. Common phenotypes are mutually accessible, translated in this figure as an almost completely connected network. As in Figures 6.1 and 6.2, phenotypes are represented as circles with radii proportional to their size. Inside each phenotype there is a complex, networked structure of genotypes, in which populations move: we have depicted a cartoon of that network inside the light yellow phenotype, highlighting in dark red a possible trajectory at the genotypic level. This trajectory implies a waiting time inside each phenotype, that is translated in phenotypic terms as stasis. Note that though the white phenotype can in principle be attained through fixation of an appropriate sequence of mutations, the time spent in the light yellow phenotype might be, in practice, much longer than that required to find the white one.

scapes picture does not include this complex dynamics, but we must keep it in mind as they will affect the evolutionary process. On another note, although any two common phenotypes are typically connected, the genotypes at the frontier can be difficult to find in the immensity of the neutral network. As a result, the appearance and fixation of adaptive mutations have a non-trivial representation in adaptive multiscapes. We can picture some trajectories in a fixed environment on the multiscape shown in Figure 6.3. Imagine that the population starts in the black phenotype, which is not

specially large or fit. We can expect some advantageous mutations to appear soon, and carry the whole population to a new phenotype. This move corresponds to an up-hill climb in the classical fitness landscape picture, in which beneficial mutations are easily found and accumulate gradually. In adaptive multiscapes, the expected dynamics are different. First, there is a variable time spent in our current phenotype, which depends on its size, as previously discussed: mutations accumulate, but no phenotypic change occurs. Second, there are several fitter phenotypes that can be accessed from the current one. The probability to jump from one to another depends on their fitness difference (Manrubia and Cuesta, 2015) and on the size of the new phenotype (Schaper and Louis, 2014). Adaptive multiscapes qualitatively capture these features in the fitness color code and in the thickness of links. Third, and at odds with the picture presented by classical fitness landscapes, the fittest phenotype need not be the one that is eventually fixed, an event that strongly depends on this phenotype's size (Schaper and Louis, 2014).

Following with our example in Figure 6.3, let us imagine the adaptive dynamics followed by a population that starts in the black phenotype. There are four fitter phenotypes reachable through this phenotype. The light yellow phenotype is very large, which makes the transition to it very likely, even though the white phenotype is fitter. Moreover, since the light yellow phenotype is fitter than the orange one, the population needs not go through this intermediate step. We need to keep in mind, however, that the stochastic nature of this process makes any feasible trajectory liable to appear in any single realization. Additionally, the time spent in each phenotype —representing stasis— will depend on the random search inside each neutral network, as previously discussed.

6.4 Empirical examples

We now discuss some specific examples to show how different dynamics can be visualized in our adaptive multiscapes framework.

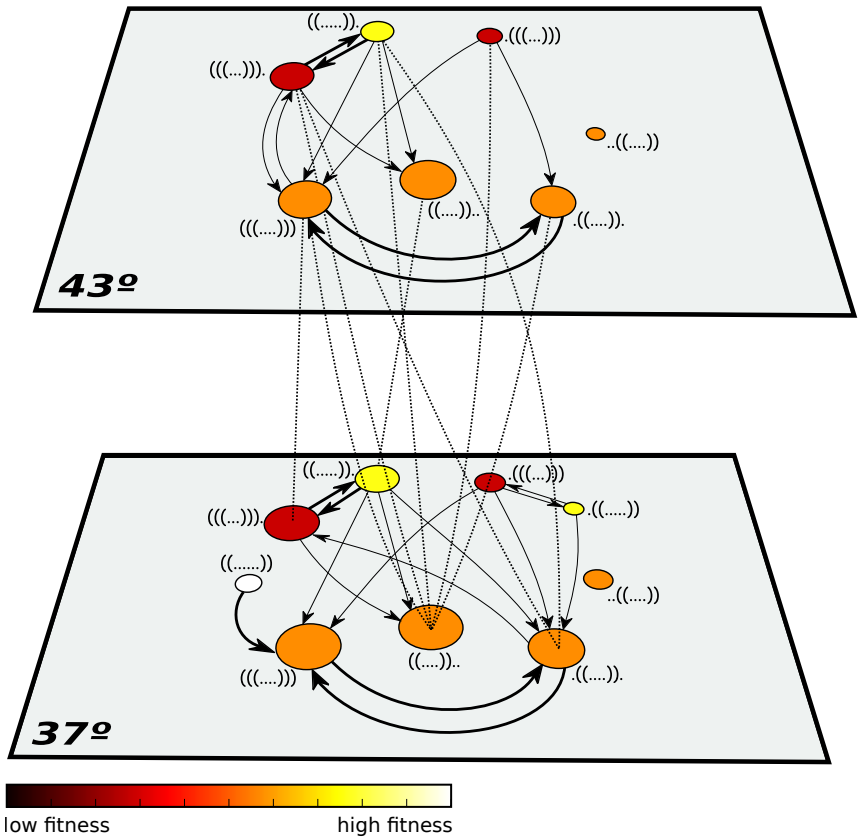


Figure 6.4: Multiscape of RNA sequences of length 10 folded at two different temperatures. There are 9 non-empty phenotypes at 37°C and 8 at 43°C, one of them having size 1 —this last phenotype is not shown in the 43°C layer. Fitness has been chosen to be proportional to the number of unpaired nucleotides in the hairpin loop, and is represented using the same color code as in previous Figures. The thickness of arrows connecting phenotypes represent the probability that a point mutation changes one phenotype into the other. Thus, thick lines represent a probability above 5%, while thin lines represent probabilities below 5%. Lower transition probabilities are not shown for clarity. When changing environments, most sequences will fold into the same structure (see Table 6.1). However, some sequences fold into a different structure at 43°C: some of these transitions between phenotypes are represented as dashed lines.

6.4.1 A synthetic quantitative example

First, we will introduce a computational example, one that can put some numbers to the intuitions we have been discussing. Consider all RNA sequences of length 10. We will, as usual, take the minimum free energy secondary structure as the phenotype. In this case, however, we will consider this phenotype as a function of two temperatures: 37°C and 43°C. The folding energy algorithm developed by the Vienna group is sensitive to temperature (Lorenz et al., 2011), so these two values will result in different genotype-to-phenotype maps. The results of this mapping are summarized in Table 6.1 and in Figure 6.4. Table 6.1 shows all non-empty phenotypes and their size at each folding temperature, as well as the fraction of neutral mutations for each phenotype at 37°C (p_{stay}^{37}) and 43°C (p_{stay}^{43}), and the probability that the phenotype is not changed when the temperature increases from 37°C to 43°C ($p_{\text{stay}}^{37\to 43}$) —we are supposing that the original temperature is 37°C. Note that $p_{\text{stay}}^{37\to 43} \neq p_{\text{stay}}^{43\to 37}$, which is not shown in the Table, and that many sequences map to the open structure: transitions to and from the open structure are also not shown.

We can take these two temperatures as two different environments. We will choose fitness to be the same in both environments, and proportional to the number of unpaired nucleotides in the hairpin loop of the secondary structure —the number of dots inside the brackets of the phenotype in the dot-bracket notation (see Table 6.1 and Figure 6.4). To slightly motivate this *ad hoc* definition, we can think of a small RNA that needs to interact with another molecule in order to perform its function: the greater the number of open nucleotides, the stronger the interaction. Using this definition, we obtain four levels of fitness (Figure 6.4).

Imagine now a population of sequences at 37°C. The fittest phenotype $((\ldots\ldots))$ is large enough, so we can assume that most of the population will be found there, if the environment has been stable for some time. However, due to the high likelihood of mutating to $(((\ldots.)))$, we will expect this second phenotype to be somewhat populated at equilibrium. A third fraction of sequences could also be found in phenotype $((\ldots..))$. —the number of individuals in this latter phenotype will depend on phenotype size, the relative transition rates between phenotypes and the difference in fitness values between phenotypes. If now the temperature increases to 43°C, the fittest

phenotype will be completely destabilized: there is only one sequence that folds into it at 43°C, it cannot be reached from any other phenotype, and mutation leads to the open structure. In practice, this means that structure $((\ldots\ldots))$ will not be present at 43°C. A population at equilibrium in 37°C will need to find the new steady state. We propose here two possible trajectories. First, if most sequences are divided between phenotypes $((\ldots\ldots))$ and $(((\ldots.)))$ at 37°C, adaptation to the fittest achievable phenotype, $((\ldots..))$. could imply traversing through less fit phenotypes when the environment changes: the population would start in its entirety at phenotype $(((\ldots.)))$ at 43°C, and this phenotype is not connected to the fittest one. A second possibility, if phenotype $((\ldots..))$. is populated at 37°C, would be immediate adaptation, given the high probability of staying in the same structure when the temperature increases (Table 6.1).

Table 6.1: Some quantitative properties of the map from RNA sequences to secondary structures at two different temperatures. The non-empty phenotypes are listed in the first column in the dot-brackets form, while the second and the third columns yield the size of the phenotype in two different environments (at two different folding temperatures, 37°C and 43°C). p^{37}_{stay} and p^{43}_{stay} are the probabilities that a point mutation does not change the phenotype at 37°C and 43°C, respectively. $p^{37\to43}_{\text{stay}}$ is the probability that a sequence folds into the same structure when the temperature increases from 37°C to 43°C.

Phenotype	Size at 37°C	Size at 43°C	p^{37}_{stay}	p^{43}_{stay}	$p^{37\to43}_{\text{stay}}$
$(((\ldots)))..$	6935	4307	0.396	0.387	0.621
$(((\ldots.)))$	7791	5149	0.386	0.366	0.658
$((\ldots.))..$	7766	5879	0.514	0.491	0.755
$((\ldots..))..$	4802	2692	0.443	0.414	0.554
$((\ldots\ldots))$	1438	1	0.409	0	0.001
$.(((\ldots)))$	2287	1542	0.384	0.366	0.671
$.((\ldots..))..$	5718	3624	0.447	0.400	0.634
$.((\ldots..))$	944	0	0.368	–	0
$..((\ldots.))$	1729	360	0.386	0.293	0.208

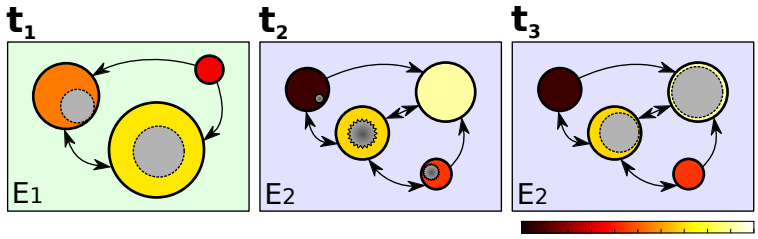


Figure 6.5: RNA virus adapting to a new environment in the framework of adaptive multiscapes. RNA virus are known for their high heterogeneity. Even at equilibrium in a given environment, they still maintain high levels of genotypic and phenotypic diversity. This diversity is represented in the first panel, at time $\mathbf{t_1}$, in environment E_1. Grey circles represent the population at equilibrium, and their radii are indicative of the degree of expansion in a neutral network. If, at a later time $\mathbf{t_2}$, the environment changes to E_2, many individuals will die, and genetic diversity is reduced: the out-of-equilibrium (maladapted) population is represented as spiked circles. The original phenotypes the virus populated in E_1 are now split into three smaller, less fit phenotypes. Through a process of mutation, the population will finally reach the new mutation-selection equilibrium at $\mathbf{t_3}$.

6.4.2 Viral populations

RNA viruses are known for their high population numbers, as well as their high genotypic and phenotypic diversity. Moreover, they are able to adapt quickly to different environments, and to escape host resistance to infection or antiviral strategies (Duarte et al., 1994). In adaptive multiscapes, viral populations will appear distributed over different phenotypes and a range of fitness values (Figure 6.5). Low fitness variants are generated constantly from high fitness genotypes as a consequence of high mutation rates, and they can become abundant in the population. In fact, if mutation rates are high enough, the fittest variant will not be the most abundant in the population (Manrubia et al., 2003). Under an environmental change, such as facing a new host (Lafforgue et al., 2011) or a new antiviral therapy (Coffin, 1995) (in Figure 6.5, the environment changes from E_1 into E_2), virus may adapt successfully through point mutations, or will be viable through functional promiscuity. These two strategies are directly translated in the visual language of adaptive multiscapes.

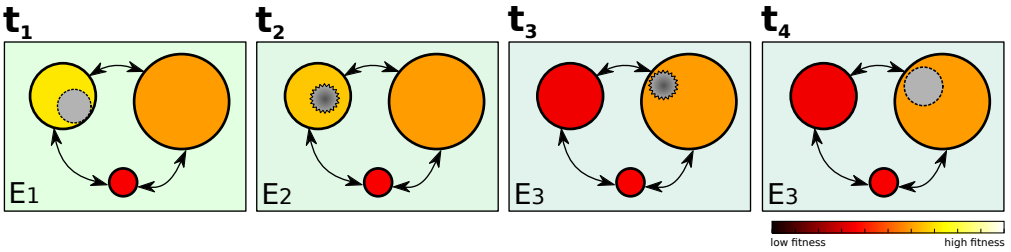


Figure 6.6: Influenza virus evolving in the framework of adaptive multiscapes. At the outset of infection season ($\mathbf{t_1}$), the virus is at equilibrium in environment E_1, populating the fittest phenotype. As time advances, however, the host population becomes immune to the virus, and the fitness of the phenotype decreases slowly ($\mathbf{t_2}$: environment changes from E_1 into E_2), until such a point at which it is lower than that of neighboring phenotypes (E_2 becomes E_3). The population then jumps to this new phenotype ($\mathbf{t_3}$), and achieves equilibrium in this new environment ($\mathbf{t_4}$). If hosts acquire immunity to this new phenotype, the picture would start again. Grey circles represent the population at equilibrium, and their radii are indicative of the degree of expansion in a neutral network. Spiked circles represent a population out-of-equilibrium.

6.4.3 Stasis, genotype network search and punctuations

Adaptive multiscapes are also able to capture the punctuated equilibria phenomenon first proposed by Eldredge and Gould (1972) and observed in molecular populations, such as influenza virus (Koelle et al., 2006). When encountering a new, fitter phenotype, the new mutant is quickly selected for, and genetic diversity decreases. The population then explores the new neutral network (thus, genetic diversity grows again), spending some time without phenotypic change (stasis). In adaptive multiscapes, this phenomenon is represented taking into account the stasis of the population during the exploration of the new phenotype —in influenza, this corresponds to infection season (Figure 6.6). As the host acquires immunity through the season, the fitness of the phenotype decreases with the number of susceptible individuals (the environment changes from E_1 to E_2 to E_3), and the probability to jump to a new phenotype increases, until the jump finally occurs (punctuation) and the virus explores the new phenotype. If hosts become immune to this new phenotype, the process will be restarted again, and we will see a new punctuating pattern.

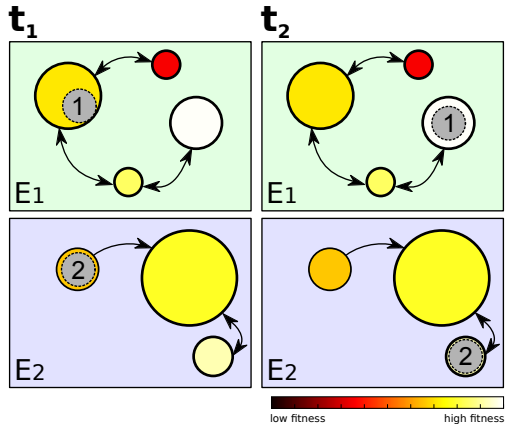


Figure 6.7: Subfunctionalization in the framework of adaptive multiscapes. Suppose a gene is subject to two different selective pressures at the same time. We can interpret this as a population that is present in two different environments E_1 and E_2. At $\mathbf{t_1}$, the population is somewhat adapted to both environments. Although in each environment there are fitter phenotypes, it is impossible for the population to increase its fitness in both environments at the same time. If gene duplication happens, this restriction disappears: it is as if the population had duplicated (grey circles 1 and 2), being able to adapt independently in each environment. As a result, in a later time $\mathbf{t_2}$, each copy of the gene will be found in a different phenotype, having optimized each function separately. Grey circles represent the population at equilibrium, and their radii are indicative of the degree of expansion in a neutral network.

6.4.4 Evolution of gene duplication

Neofunctionalization and subfunctionalization are two of the mechanisms proposed to explain the persistence of gene duplications in time (Innan and Kondrashov, 2010). In the first case, the duplicated gene can accumulate mutations in a neutral way —while the original copy still performs its function. The exploration of genotype space by the duplicated gene is enhanced —there are almost no restrictions— until it finds a new function. Once this happens, the duplicated version will be optimized for the new function. In the case of subfunctionalization, it is understood that the original gene performs more than one function (Figure 6.7): it is as if the population was present in two environments E_1 and E_2 at the same time. Once the gene

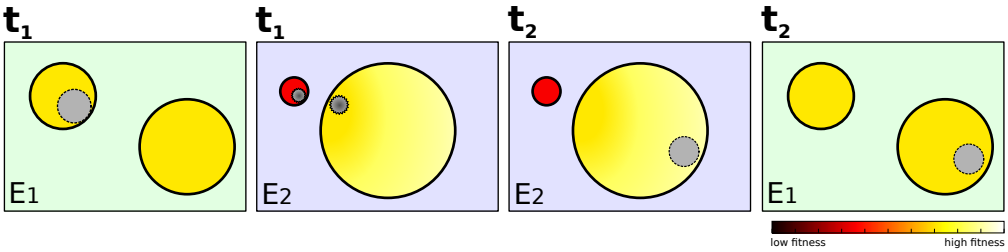


Figure 6.8: Genetic assimilation in the framework of adaptive multiscapes. A population at equilibrium in environment E_1 is phenotypically homogeneous in that environment, but not in E_2, where it expresses two different phenotypes ($\mathbf{t_1}$). If the environment changes to E_2, the population is not an equilibrium and will try to adapt to the new situation. Note that the larger phenotype in E_2 is not homogeneous in terms of fitness. At a later time $\mathbf{t_2}$, the population has reached a new equilibrium in E_2, but the phenotype it expresses is not the initial one anymore. If, going back to E_1, this new phenotype is the same as the one in E_2, then the phenomenon is called genetic assimilation: a phenotype that was expressed only promiscuously in a different environment E_2 is now expressed in the initial environment E_1. Grey circles represent the population at equilibrium, and their radii are indicative of the degree of expansion in a neutral network. Spiked circles represent a population out-of-equilibrium.

is copied, the two copies will become free to specialize in each environment independently. As a result, the two copies will diverge and end up in different phenotypes.

6.4.5 Waddington's genetic assimilation

Genetic assimilation is a phenomenon proposed and experimentally observed by Conrad Waddington (1953). Starting with a population of fruit flies, he observed that some of them expressed aberrant phenotypes when exposed to abnormal environmental conditions at the larval stage. He selected the flies that expressed this strange phenotype and, after a few generations, found that the offspring expressed the aberrant phenotype in normal conditions —the phenotype had become "assimilated" in the genotype. This observation can be understood in terms of adaptive multiscapes in the following way (Figure 6.8): a diverse population starts at a given phenotype in an initial environment E_1, and functional promiscuity causes it to

express different phenotypes in a different environment E_2 —the abnormal environmental conditions that larvae were subjected to in Waddington's experiments. Forced to evolve in this new environment E_2, the population diffuses through the network of genotypes associated to the new phenotype —note that there can be fitness differences inside this complex multicellular phenotype—, and reaches an area which, when the environment reverts to its original condition, does not express the original phenotype anymore. The aberrant phenotype is thus expressed in the original environment instead of the initial phenotype (Figure 6.8).

6.5 Summary

In this chapter we have presented a new metaphor for the adaptive process that integrates Wright's fitness landscape with important features of molecular evolution unknown in Wright's time, namely the existence of neutral networks, the uneven distribution of phenotype sizes and the asymmetric accessibility of phenotypes, and functional promiscuity. As a result, many complex features of molecular evolution can be visually captured in our picture. We have also rephrased specific examples under the light of our framework, hoping to illustrate the potential of this metaphor.

Adaptive multiscapes are not without limitations. We have intended them to capture the dynamics of molecular evolution, and are therefore not suited to describe the evolution of complex organisms, in which regulatory and developmental process interact in complicated ways with the environment to define the phenotype. Additionally, frequency-dependent selection, in which fitness values depend on the composition of the population, cannot be easily included in our framework. Finally, there are cases in which the high dimensionality of genotype space will not be relevant for the evolutionary process, and in those cases the fitness landscape metaphor could still be useful instead of adaptive multiscapes.

We have kept the discussion of evolutionary dynamics at a qualitative level in this chapter, but many of these features can be expressed quantitatively, as we have seen in Chapters 3 and 4. The topological properties of neutral networks will have an effect in the ability to reach new phenotypes (Wagner, 2011), as well as in the time required to reach equilibrium (Aguirre et al., 2009) and in the waiting time inside each phenotype, be-

fore moving to a new one (Manrubia and Cuesta, 2015). The description of these properties is still a work in progress, one which this thesis (humbly) intends to contribute to.

7

Conclusions and future work

> Adrian Veidt: "I did the right thing, didn't I? It all worked out in the end".
> Dr. Manhattan: " 'In the end'? Nothing ends, Adrian. Nothing ever ends".
>
> *Alan Moore*
> Watchmen # 12 (1987)

This thesis set out to contribute to the growing body of knowledge pertaining models of the genotype-phenotype map. In the process, we proposed and studied a new computational model —toyLIFE— and a new metaphor for molecular evolution —adaptive multiscapes. We also studied functional promiscuity and the evolutionary dynamics of shifting environments. In this chapter we will summarize the most important conclusions of this thesis, and we will outline some research lines for future work.

7.1 toyLIFE

The first result of this thesis was the definition of toyLIFE, a simplified and unrealistic model of metabolism that incorporated toy versions of genes, proteins and regulation as well as metabolic laws. Molecules in toyLIFE interact between each other following the laws of the HP protein folding model, which endows toyLIFE with a simplified chemistry. From these laws, we saw how something reminiscent of cell-like behavior emerged, with complex regulatory and metabolic networks that grew in complexity as the genome increased.

toyLIFE was born to confirm the intuition that a multi-level genotype-phenotype map would differ significantly from the previous models studied in the literature, such as RNA, proteins, GRNs or metabolic networks. All of these models either disregarded cellular context when assigning phenotype and function (RNA and proteins) or omitted genome dynamics, by defining their genotypes from high-level abstractions (GRNs and metabolic networks). Surprisingly, most of the results we found when studying the genotype-phenotype map defined in toyLIFE were very similar to those observed in those models (see Chapter 3). There was strong degeneracy in the map, with many genotypes mapping into the same phenotype. This degeneracy translated into the existence of neutral networks, that spanned genotype space as soon as the genotype contained more than two genes. There was also a strong asymmetry in the size distribution of phenotypes: most phenotypes were rare, while a few of them covered most genotypes. Moreover, most common phenotypes were easily accessed from each other, and the *shape space covering* property described for RNA was also observed here.

Interestingly, while HP protein folding model does not show the shape space covering property, toyLIFE, which is built on it, does —and on more than one level, too: we observed it for the metabolic phenotype in Chapter 3 and for the regulatory, pattern-generating one in Chapter 5. This fact is, again, another confirmation of Wagner's observation that robustness and evolvability are positively correlated (Wagner, 2011). Adding more layers of complexity to the genotype-phenotype map, robustness must necessarily increase —many combinations of proteins will produce the same expression pattern, and different proteins with different expression patterns will

have the same metabolic phenotype. With robustness increased, the number of contacts between phenotypes increases, and with that accessibility grows as well. The increase of robustness with the added complexity is also related to Waddington's concept of canalization (Waddington, 1942). In his view of development, Waddington understood that many different genes interacted between each other and buffered the effect of environmental changes, so that the embryo could grow healthily. As complexity grows, however, this canalization becomes more and more entrenched, and small changes are fatal for development. This restraining character of complexity has also been observed by Erwin and Davidson (Davidson and Erwin, 2006; Erwin and Davidson, 2009). We have yet to study if adding more and more genes to toyLIFE genotypes will result in constrained genotypes such as those observed in multicellular animals, for example.

We also observed in toyLIFE the linear relationship between phenotypic robustness and the logarithm of phenotype size as well as the log-normal distribution of phenotypes, both observations already described for RNA secondary structure (Aguirre et al., 2011; Dingle et al., 2015). The fact that these two models, which have almost nothing in common, present these two features points to a fundamental property of models of the genotype-phenotype map. Manrubia and Cuesta (2017) have proposed a combinatorial argument that would explain the appearance of the log-normal distribution for simple models of the map. We have yet to study if toyLIFE, as well as other maps such as RNA, fulfill the conditions exposed in that paper. It is our intuition that the linear relationship between phenotypic robustness and phenotype size is a result of those combinatorial arguments, and in Chapter 3 we gave some heuristic arguments that were meant to explain this relationship in toyLIFE. We will need to study how general they are, in connection to other computational models. This research program may seem abstract at the outset. However, if we can find general arguments that link all these properties together, and if we can extrapolate them to real systems, we could make interesting evolutionary predictions for real systems —whose study is out of our computational and modelling possibilities at the moment.

7.1.1 (Future work) Extensions to toyLIFE

We need not point out that toyLIFE has many more possibilities than those studied in this thesis. The first example are the pattern-formation, multicellular regulatory genomes, of which we have barely scratched the surface in Chapter 5. A full characterization of that genotype-phenotype map, plus a comparison with the results shown in Chapter 3, would help to complete our knowledge of this model. Besides, the complex and intricate patterns generated by toyLIFE present many interesting features that will need further attention. Different signaling mechanisms could also be implemented in toyLIFE, enabling us to explore relevant questions related to development. Moreover, the extension of this phenotype to a two-dimensional tissue is fairly easy, and we could study the evolution of two-dimensional patterns. Because toyLIFE includes many complexity levels underlying this phenotype, it could yield insights into how complex phenotypes are built.

But toyLIFE's potential does not stop there. Gene duplications are easy to implement in this system, and the exploration of a genome space in which duplications and deletions are valid mutational moves would be an interesting generalization of the hypercubic spaces studied so far in the literature.

Additional genomic questions could be addressed with toyLIFE. For instance, we could understand toyGenes as fragments of a larger genome in which they are embedded, as real genes are. We would need to include some kind of signal for the initiation of translation, so that the polymerase knows what is a gene and what is not. This would lead to interesting dynamics with overlapping genes. This dynamics could be greatly enriched by the addition of transposons or viruses, genetic elements able to insert themselves in the genome and disrupt its functioning. toyLIFE would allow us to study not only that disruption, but its effect on every genotype, and the resulting consequences on robustness and evolvability.

Studying genomic structure would also allow us to study the evolution of *junk* genes in depth. As we saw in Chapter 3, toyLIFE genotypes tend to become rich with junk genes as they grow. Most of the genes that we add to a genotype will not affect their metabolic function —however, they increase the genotype's robustness and evolvability, so they are not strictly

inert. We wonder what would happen if those genes were embedded in a genomic structure that could grow and shrink due to duplications and deletions. It would be interesting to study the dynamics of junk genes in that scenario, observing if robustness and evolvability are increased or if, on the contrary, their effect is mostly deleterious.

Alternatively, it would be interesting to study the dynamics of regulation with changing promoter regions. Eukaryotic genomes are known to have complicated regulatory signals, in which many different transcription factors bind to different parts of the genome before and after the genes (Ptashne and Gann, 2002; Alberts et al., 2014). We could introduce some interaction rules that would allow for the functioning of longer promoters. For instance, we could say that two toyProteins can only bind to adjacent promoter sides if they form a toyDimer. With such a mechanism, we could introduce mutations that make the promoter region grow or shrink, and study their effect on metabolism and regulation. Longer promoters lead to more complex regulatory dynamics, in which many different expression routines can be carried out simultaneously. At the same time, the longer the promoter, the greater the probability of mutations, thus creating the need for more robustness in promoters. The final, optimal length of a promoter will be a trade-off between flexibility and robustness. Studying that trade-off would be an interesting question.

On a different note, we could also include the existence of mistranslating proteinns and study their effect on the regulatory and metabolic phenotype. Mistranslation is a phenotypic mutation in which a protein is synthesized with an incorrect amino acid sequence, due to errors in the translation process (Bratulic et al., 2015). With toyLIFE we can exhaustively study the effect of varying levels of mistranslation on metabolism, and understand how robustness to mistranslation can evolve in such a simple system.

7.1.2 (Future work) Ecology in toyLIFE

All extensions to toyLIFE discussed in the previous paragraphs mostly deal with unicellular genotypes. And, when they deal with multicellular organisms, as in the case of regulatory patterns, we are always considering one common genotype for all cells. toyLIFE can be further extended, however, to include ecological interactions. In order to do that, we will need to de-

vise some kind of reproduction and death mechanisms, introducing some elementary energetics into the model. Suppose, for example, that a cell can divide once it has "eaten" so and so molecules of toySugar, or so and so toyProteins. And that it will die if it has no toySugar molecules left inside itself. The possibilities are infinite, so this step of the process will need much attention and care.

Whatever the final details of the extended ecological toyLIFE model, it is not difficult to imagine that once we have independent cells that reproduce and die according to some rules, ecological interactions will appear naturally. In an environment with finite resources —in the form of toyMetabolites or toyProteins— cells will compete for these, and natural selection will take place. We could also introduce some mechanisms for predation, or infection, thus endowing the system with all ingredients needed for complex ecosystems to arise.

With all those ingredients, we could investigate the interplay between ecology and evolution, a highly interesting interface sadly neglected by traditional evolutionary and ecological theory. Evolutionary theory has proposed several models in which population size is considered to be constant, while ecologists have devised multitude of models that study the variation of population size, but without any evolution. There are exceptions to this trend, however, with one example being the adaptive dynamics models, ecological models in which the parameters evolve through time (Metz et al., 1995; Dieckmann, 1997).

In this sense, toyLIFE could become another artificial life model, such as the famous Avida (Ofria and Wilke, 2004). Avida contains organisms whose genomes are computer instructions. They live in an environment full of strings that the organisms "eat" as input, and their fitness is measured by their ability to generate certain outputs from given inputs. These organisms are able to self-replicate as part of their computer instructions. The advantage of our ecological toyLIFE model over Avida is that fitness would be born out of the environment, instead of being defined *ad hoc*. If an organism is able to "eat" faster than others, then it will reproduce faster, and will eventually dominate the population. In the process, it will mutate and improve, and we can study what happens when the environment changes and many other scenarios. The point, however, is that we do not define *a priori* which genotype will be the fittest. Each realization of

the process would be different, depending on the initial conditions and the environmental stimuli. Ecological toyLIFE would therefore become a toy laboratory where we could study many aspects of the evolutionary process, performing computational experiments to test our hypotheses.

The implementation of ecology in toyLIFE would take some time, and deeper programming skills than those we have needed so far. However, we hope to be able to develop this extension in the future, as it is full of promises.

7.1.3 (Future work) toyLIFE as a didactical tool

Our usual ways of thinking are full of biases, and our reasoning is derailed by them. The scientific method is designed to help us overcome these biases and, while the scientific endeavor is far from perfect, I believe it is the best tool we have to understand the world. But teaching the scientific method is not always easy. From my experience as a biology undergraduate, I have seen how many of my classmates left University without a full and correct grasp of the scientific method: their approach to science was full of dogma and belief, and they did not understand the power of scientific thinking.

toyLIFE can become an interesting tool to teach about the scientific method, either to teenagers or to undergraduate students. There are many ways in which this tool can be implemented, of course. One particular choice could be as follows: we could design experiments to obtain information about toyLIFE —just like we do in real biological systems. The teacher could then run the experiments computationally and give the students the results. The students would have to analyze those results and come up with theories able to explain them. The interesting part is that students should be able to propose new experiments that would confirm or disprove their theories, and learn how to adapt their understanding as they perform subsequent experiments and obtain new data. toyLIFE seems a good choice for this kind of teaching system, as we do know everything about its rules —indeed, we designed them ourselves. toyLIFE shows a complex behavior, enough to generate interesting datasets that can puzzle students over and over, helping them get a sense of how science really works.

7.2 Functional promiscuity

In Chapter 4 we studied the prevalence of functional promiscuity in computational models of the genotype-phenotype map. In that chapter, we saw that promiscuity is the norm, rather than the exception. These results prompt us to rethink our understanding of biology as a neatly functioning machine. This understanding has recently started to grow in the scientific community (Daniels et al., 2008; Tawfik, 2010), but we need to take it further. In order to fully understand the complexities of evolution, we need to start including functional promiscuity in our models.

One way to further explore this promiscuity in $\mathtt{toy}$LIFE would be to relax the disambiguity rules, allowing for toyProteins that can fold in more than one structure. We could start by considering those toyProteins we discarded in Chapter 2 because they folded in two different structures with the same energy and perimeter. These promiscuous toyProteins would be able to perform more than one function —but because the folding energy of the two folds is the same, their function would need to be stochastic. As a result, the desired function would be performed only sometimes, resulting in a fitness cost. An interesting question would be to explore under which conditions this kind of toyProteins are selected for: the presence of environmental changes or stochasticity seem good candidates, for instance. We could expand this initial definition of promiscuity by allowing all toyProteins to fold stochastically into different structures according to their folding energy, mimicking what real RNA and proteins do in real cells. This expansion would allow us to study the conditions needed to select for robustness or promiscuity, in a system in which the phenotype is given by a higher-level function, such as metabolism or regulation. This extension to $\mathtt{toy}$LIFE would also allow us to explore the consequences of low-level promiscuity on metabolism and regulation: we could observe how messy they become —or if they become messy at all. It would be interesting to analyze these new networks, as well as the appearance of buffering mechanisms, among many other features.

7.3 Dynamics of shifting environments

One of the most interesting results of this thesis came from studying the evolutionary dynamics of shifting environments. A naïve approach to this problem would suggest that, if we want to eliminate a population, subjecting it to very fast environmental changes is the best option. However, this is not what we observe in our simulations. Our results show that there is an optimal frequency of change that minimizes the time to extinction of the population. However, there are many parameters involved in the simulations, from the population's mutation and death rates, to the topology of genotype space or the relationship between fitness values in both environments. We need to keep exploring these results, devising mathematical models that give us some insight into the underlying dynamics. Some simplifications will need to be made, however, as the complexity of the model studied in Chapter 4 is already too high to allow analytical study.

This particular line of research is especially interesting because it ties in with real problems in medicine, namely the treatment of bacterial diseases. Bacteria are known to evolve resistance to the antibiotics we use to treat them, posing a serious threat for public health, once all of our antibiotics are useless. Our theoretical work, as well as experimental results (Fuentes-Hernandez et al., 2015), suggest that all hope is not lost, and that we can re-use our antibiotics in intelligent ways to end with bacterial infections. Previous attempts to combine antibiotics had used them simultaneously, which we could interpret as very fast environmental changes in our model. As we have seen, if the population is not completely extinct at the beginning of this shock, then it will be able to survive and thrive in the new, constant environment. However, alternating between antibiotics at a slower rate maximizes the probabilities of extinction. We need to continue exploring this question, both theoretically and experimentally. Other antibacterial treatments are being developed in recent years, such as phage therapy and antimicrobial peptides, but we believe antibiotic cycling to be an exciting and promising avenue, which remains largely unexplored.

7.4 Adaptive multiscapes

Chapter 6 worked as a kind of summary of all the insights gained throughout the thesis, combined in a new metaphor for molecular evolution: adaptive multiscapes. This framework intends to update the fitness landscape metaphor proposed by Sewall Wright in the 30s (Wright, 1932). Adaptive multiscapes include many features that we have learned from computational studies of the genotype-phenotype map, and that have been discussed throughout the thesis. The existence of neutral networks, the asymmetry in phenotype sizes —and the concomitant asymmetry in phenotype accessibility— and the presence of functional promiscuity all alter the original fitness landscape picture.

Our qualitative presentation of adaptive multiscapes has allowed us to rephrase some evolutionary phenomena in terms of our new metaphor, showing its intuitive potential. Adaptive multiscapes, however, also allow for quantitative exploration of evolutionary phenomena, and it is our hope that we can —at some point in the future— develop a mathematical framework that gathers all this intuitions and yields correct predictions about evolutionary dynamics.

Until that day arrives, however, we hope the wider community of evolutionary biologists embraces this new framework, discussing and improving it, so that our intuitions on the evolutionary process can be refined and polished.

A

Appendix

A.1 Prediction of promiscuity for GRNs using analytic combinatorics

Boolean networks of g genes are defined by Boolean functions $F : \{0,1\}^g \mapsto \{0,1\}^g$. By interpreting each sequence of g bits as the binary representation of an nonnegative integer, Boolean functions are mappings $f : N \mapsto N$, where $N = \{0,1,\ldots,2^g-1\}$.

Mappings of this sort can be represented using a graph. Starting from $x_0 \in N$, the sequence $x_{n+1} = f(x_n)$ eventually ends in a fixed point or in a cycle. If we join with a link x_n with x_{n+1} for all n and all initial points $x_0 \in N$, we end up with a labelled graph $\mathcal{G}_f$. Function f can be reconstructed from the graph, so there is a one-to-one mapping between Boolean functions and graphs.

The nature of these graphs can be identified by construction: they are made of sets of connected graphs, each of which consists of a set of Cayley trees (arbitrary number of branches, branch order being irrelevant) rooted in a cycle. Accordingly, using the symbolic method described in Flajolet and Sedgewick (2009) we can count f_n, the total number of mappings f :

$N \mapsto N$, by counting the number of graphs of $n = |N| = 2^g$ nodes. This can be achieved by directly obtaining the exponential generating function (EGF)

$$F(z) = \sum_{n=0}^{\infty} \frac{f_n}{n!} z^n. \tag{A.1}$$

If $\mathcal{F}$ is the combinatorial class of graphs that represent mappings, then

$$\mathcal{F} = \mathsf{SET}(\mathcal{K}), \qquad \mathcal{K} = \mathsf{CYC}(\mathcal{T}), \qquad \mathcal{T} = \mathcal{Z} * \mathsf{SET}(\mathcal{T}), \tag{A.2}$$

where $\mathcal{K}$ is the combinatorial subclass of such connected graphs, and $\mathcal{T}$ that of Cayley trees. Thus (Flajolet and Sedgewick, 2009, sec. II.5.2)

$$F(z) = e^{K(z)}, \qquad K(z) = -\log[1 - T(z)], \qquad T(z) = z e^{T(z)}, \tag{A.3}$$

where $K(z)$ and $T(z)$ are the corresponding EGFs of $\mathcal{K}$ and $\mathcal{T}$. Therefore $F(z) = [1 - T(z)]^{-1}$ and, using Lagrange's inversion formula (Flajolet and Sedgewick, 2009, sec. A.6), $f_n = n![z^n]F(z) = n^n$.

If we want to count $f_{n,k}$, the number of different graphs with k different cycles (a fixed point counts as a cycle of length one) we need to introduce the bivariate EGF

$$F(z,u) = \sum_{n=0}^{\infty} \frac{f_n(u)}{n!} z^n, \qquad f_n(u) = \sum_{k=0}^{\infty} f_{n,k} u^k. \tag{A.4}$$

Using the symbolic method

$$F(z,u) = e^{uK(z)} = [1 - T(z)]^{-u}. \tag{A.5}$$

We can estimate the asymptotic behavior of $f_{n,k}$ when $n \to \infty$. To that purpose it is convenient to use the asymptotic estimate for $z \to \left(e^{-1}\right)^-$

$$T(z) = 1 - \sqrt{2}(1 - ez)^{1/2} + O(1 - ez), \tag{A.6}$$

which leads to

$$F(z,u) = 2^{-u/2}(1 - ez)^{-u/2} + O\left((1 - ez)^{(1-u)/2}\right). \tag{A.7}$$

Using Darboux's theorem (reference in Flajolet and Sedgewick (2009)), for $n \to \infty$

$$f_n(u) = n! \frac{2^{-u/2}}{\Gamma(u/2)} n^{-1+u/2} e^n \left[1 + O\left(n^{-1/2}\right)\right], \tag{A.8}$$

which, given that $n! \sim \sqrt{2\pi n} n^n e^{-n}$ and that $\Gamma(u/2) = (2/u)\Gamma(1+u/2)$, can be rewritten as

$$f_n(u) = n^n \frac{\sqrt{\pi}}{2\Gamma(1+u/2)} u e^{\lambda_n(u-1)} \left[1 + O\left(n^{-1/2}\right)\right], \qquad \lambda_n = \frac{1}{2}\log\left(\frac{n}{2}\right). \tag{A.9}$$

Clearly in this expression $f_n(u) = f_n \mathbb{E}(u^X)$, where X is a random variable such that $\Pr\{X = k\} = f_{n,k}/f_n$. Now, $\mathbb{E}(u^X)$ is a product of two terms: one is $ue^{\lambda_n(u-1)}$, which corresponds to the shifted Poisson process

$$\Pr\{X_1 = k\} = \begin{cases} 0, & k = 0, \\ e^{-\lambda_n} \dfrac{\lambda_n^{k-1}}{(k-1)!}, & k > 0; \end{cases} \tag{A.10}$$

the second one is $\sqrt{\pi}/2\Gamma(1+u/2)$. To figure out the nature of this process we notice that $\Pr\{X_2 = 0\} = \sqrt{\pi}/2\Gamma(1) = \sqrt{\pi}/2 \approx 0.886$. In other words, the random variable X describing our process is the sum $X = X_1 + X_2$, where X_1 is the shifted Poisson process described above and $X_2 = 0$ with nearly 89% probability. Thus, to a very good approximation our process is given by the shifted Poisson.[1]

A.2 Obtaining the total number of phenotypes for GRNs

A phenotype in a GRN is any cycle that can be obtained with a subset of $N = \{0, 1, \ldots, 2^g - 1\}$. The length of the cycle, j, can go from $j = 1$ — for point attractors— to $j = 2^g = n$ —for cycles that traverse all possible

[1] Rigorously speaking, $\sqrt{\pi}/2\Gamma(1+u/2)$ does not correspond to a true probability distribution because its expansion in powers of u has negative coefficients (e.g., those of u^2 and u^5). They are small though, so we can take the function as a reasonable approximation to the expectation $\mathbb{E}(u^{X_2})$ of a genuine process, whose probability is mostly concentrated at $X_2 = 0$.

states. For any given length j, we have to choose from the set of n states, and there are $(j-1)!$ cycles with a labelled set of length j. Thus the total number of phenotypes for a given n, F_n, is immediately given by

$$F_n = \sum_{j=1}^{n} \binom{n}{j} (j-1)!. \tag{A.11}$$

In order to get an asymptotic expression for F_n, we use EGFs again:

$$A(z) = \sum_{n=1}^{\infty} F_n \frac{z^n}{n!} = \sum_{n=1}^{\infty} \sum_{j=1}^{n} \frac{(j-1)!}{j!(n-j)!} z^n = \sum_{j=1}^{\infty} \frac{z^j}{j} \sum_{n=j}^{\infty} \frac{z^{n-j}}{(n-j)!} = -e^z \log(1-z), \tag{A.12}$$

in which we have changed the summation order. Using the results from (Flajolet and Sedgewick, 2009, p.449), we obtain that

$$F_n = n![z^n]A(z) = e(n-1)! \left[1 + O\left((\log n)^{-2}\right)\right]. \tag{A.13}$$

Publications

The original content of this thesis appears (or will appear) in the following papers:

Chapter 2

toyLIFE: a computational framework to study the multi-level organization of the genotype-phenotype map, C. F. Arias, **P. Catalán**, S. C. Manrubia and J. A. Cuesta, *Sci. Rep.* **4**, 7549 (2014).

Chapter 3

Robustness and evolvability in a metabolic genotype-phenotype map, **P. Catalán**, C. F. Arias, S. C. Manrubia and J. A. Cuesta, in preparation.

Chapter 4

Functional promiscuity in computational models of the genotype-phenotype map, **P. Catalán**, S. C. Manrubia and J. A. Cuesta, in preparation.

Killing bugs the smart way: using evolutionary theory to design new antibiotic therapies, **P. Catalán**, S. C. Manrubia and J. A. Cuesta, in preparation.

Chapter 5

Complex regulatory spatio-temporal patterns in toyLIFE, **P. Catalán**, S. C. Manrubia and J. A. Cuesta, in preparation.

Chapter 6

Adaptive multiscapes: an up-to-date metaphor to visualize molecular evolution, **P. Catalán**, C. F. Arias, J. A. Cuesta and S. C. Manrubia, accepted in *Biol. Dir.* (2017).

Bibliography

Aguirre, J., J. M. Buldú, M. Stich, and S. C. Manrubia (2011). Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLoS ONE 6*, e26324.

Aguirre, J., J. M. Buldú, and S. C. Manrubia (2009). Evolutionary dynamics on networks of selectively neutral genotypes: Effects of topology and sequence stability. *Phys. Rev. E 80*, 066112.

Aharoni, A., L. Gaidukov, O. Khersonsky, S. M. Gould, C. Roodveldt, and D. S. Tawfik (2005). The'evolvability'of promiscuous protein functions. *Nat. Genet. 37*, 73–76.

Alberch, P. (1991). From genes to phenotype: dynamical systems and evolvability. *Genetica 84*, 5–11.

Albert, R. and H. G. Othmer (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J. Theor. Biol. 223*, 1–18.

Alberts, B., A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter (2014). *Molecular Biology of the Cell, 6th edition*. Garland Science.

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC press.

Amitai, G., R. D. Gupta, and D. S. Tawfik (2007). Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J. 1*, 67–78.

Ancel, L. W. and W. Fontana (2000). Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool. 288*, 242–283.

Arias, C. F., P. Catalán, S. Manrubia, and J. A. Cuesta (2014). toyLIFE: a computational framework to study the multi-level organisation of the genotype-phenotype map. *Sci. Rep. 4*, 7549.

Babajide, A., I. L. Hofacker, M. J. Sippl, and P. F. Stadler (1997). Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold. Des. 2*, 261–269.

Barve, A. and A. Wagner (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature 500*, 203–206.

Bastolla, U., H. E. Roman, and M. Vendruscolo (1999). Neutral evolution of model proteins: diffusion in sequence space and overdispersion. *J. Theor. Biol. 200*, 49–64.

Berger, B. and T. Leighton (1998). Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J. Comput. Biol. 5*, 27–40.

Bloom, J. D., A. Raval, and C. O. Wilke (2007). Thermodynamics of neutral protein evolution. *Genetics 175*, 255–266.

Bornberg-Bauer, E. (1997). How are model protein structures distributed in sequence space? *Biophys. J. 73*, 2393.

Bowler, P. J. (1989). *Evolution: the History of an Idea*. Univ of California Press.

Bratulic, S., F. Gerber, and A. Wagner (2015). Mistranslation drives the evolution of robustness in tem-1 β-lactamase. *Proc. Natl. Acad. Sci. USA 112*, 12758–12763.

Byun, Y. and K. Han (2006). Pseudoviewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Res. 34*, W416–W422.

Byun, Y. and K. Han (2009). Pseudoviewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinform. 25*, 1435–1437.

Cervera, H., J. Lalić, and S. F. Elena (2016). Effect of host species on topography of the fitness landscape for a plant rna virus. *J. Virol. 90*, 10160–10169.

Cheng, D., H. Qi, and Z. Li (2011). *Analysis and Control of Boolean Networks*. Springer.

Ciliberti, S., O. C. Martin, and A. Wagner (2007a). Innovation and robustness in complex regulatory gene networks. *Proc. Natl. Acad. Sci. USA 104*, 13591–13596.

Ciliberti, S., O. C. Martin, and A. Wagner (2007b). Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput. Biol. 3*, e15.

Coffin, J. M. (1995). Hiv population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science 267*, 483.

Corrales-Guerrero, L., E. Flores, and A. Herrero (2014). Relationships between the ABC-exporter HetC and peptides that regulate the spatiotemporal pattern of heterocyst distribution in *Anabaena*. *PloS ONE 9*, e104571.

Cotterell, J. and J. Sharpe (2010). An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. *Mol. Sys. Biol. 6*, 425.

Daniels, B. C., Y.-J. Chen, J. P. Sethna, R. N. Gutenkunst, and C. R. Myers (2008). Sloppiness, robustness, and evolvability in systems biology. *Curr. Opin. Biotech. 19*, 389–395.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (1st ed.). London: John Murray.

Davidson, E. H. and D. H. Erwin (2006). Gene regulatory networks and the evolution of animal body plans. *Science 311*, 796–800.

De Visser, J. A. G. and J. Krug (2014). Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet. 15*, 480–490.

De Vos, M. G., A. Dawid, V. Sunderlikova, and S. J. Tans (2015). Breaking evolutionary constraint with a tradeoff ratchet. *Proc. Natl. Acad. Sci. USA 112*, 14906–14911.

Dieckmann, U. (1997). Can adaptive dynamics invade? *Trends Ecol. Evol. 12*, 128–131.

Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry 24*, 1501–1509.

Dingle, K., S. Schaper, and A. A. Louis (2015). The structure of the genotype–phenotype map strongly constrains the evolution of non-coding RNA. *Interface focus 5*, 20150053.

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. Columbia University Press.

Duarte, E. A., I. S. Novella, S. Ledesma, D. K. Clarke, A. Moya, S. F. Elena, E. Domingo, and J. J. Holland (1994). Subclonal components of consensus fitness in an RNA virus clone. *J. Virol. 68*, 4295–4301.

Eaton, W. A. and J. Hofrichter (1990). Sickle cell hemoglobin polymerization. *Adv. Prot. Chem. 40*, 63–279.

Eldredge, N. and S. J. Gould (1972). Punctuated equilibria: an alternative to phyletic gradualism. In T. J. M. Schopf (Ed.), *Models in Paleobiology*, pp. 82–115. San Francisco: Freeman Cooper.

Elena, S. F. and R. E. Lenski (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet. 4*, 457–469.

Erwin, D. H. and E. H. Davidson (2009). The evolution of hierarchical gene regulatory networks. *Nat. Rev. Genet. 10*, 141–148.

Espinosa-Soto, C., O. C. Martin, and A. Wagner (2011). Phenotypic plasticity can facilitate adaptive evolution in gene regulatory circuits. *BMC Evol. Biol. 11*, 5.

Espinosa-Soto, C., P. Padilla-Longoria, and E. R. Alvarez-Buylla (2004). A gene regulatory network model for cell-fate determination during *Arabidopsis thaliana* flower development that is robust and recovers experimental gene expression profiles. *Plant Cell 16*, 2923–2939.

Eyre-Walker, A. and P. D. Keightley (2007). The distribution of fitness effects of new mutations. *Nat. Rev. Genet. 8*, 610–618.

Ferrada, E. and A. Wagner (2012). A comparison of genotype-phenotype maps for RNA and proteins. *Biophys. J. 102*, 1916–1925.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford University Press.

Fishman, N. (2006). Antimicrobial stewardship. *A. J. Infect. Control 34*, S55–S63.

Flajolet, P. and R. Sedgewick (2009). *Analytic Combinatorics*. Cambridge: Cambridge University Press.

Fontana, W. (2006). The topology of the possible. In *Understanding Change*, pp. 67–84. Springer.

Fontana, W. and P. Schuster (1998). Continuity in evolution: on the nature of transitions. *Science 280*, 1451–1455.

Fuentes-Hernandez, A., J. Plucain, F. Gori, R. Pena-Miller, C. Reding, G. Jansen, H. Schulenburg, I. Gudelj, and R. Beardmore (2015). Using a sequential regimen to eliminate bacteria at sublethal antibiotic dosages. *PLoS Biol. 13*, e1002104.

Gama-Castro, S., H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón, et al. (2015). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.*, gkv1156.

Garcia-Ojalvo, J. (2011). Physical approaches to the dynamics of genetic circuits: a tutorial. *Contemp. Phys. 52*, 439–464.

Gavrilets, S. (1997). Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol. 12*, 307–312.

Gavrilets, S. and J. Gravner (1997). Percolation on the fitness hypercube and the evolution of reproductive isolation. *J. Theor. Biol. 184*, 51–64.

Gillespie, J. H. (1991). *The Causes of Molecular Evolution*. Oxford University Press.

Goldberg, A. D., C. D. Allis, and E. Bernstein (2007). Epigenetics: a landscape takes shape. *Cell 128*, 635–638.

Gould, S. J. and N. Eldredge (1977). Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology 3*, 115–151.

Greenbury, S. and S. Ahnert (2015). The organization of biological sequences into constrained and unconstrained parts determines fundamental properties of genotype–phenotype maps. *J. R. Soc. Interface 12*, 20150724.

Greenbury, S. F., I. G. Johnston, A. A. Louis, and S. E. Ahnert (2014). A tractable genotype-phenotype map modelling the self-assembly of protein quaternary structure. *J. R. Soc. Interface 6*, 20140249.

Grüner, W., R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster (1996a). Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Monatshefte f. Chemie 127*, 355–374.

Grüner, W., R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster (1996b). Analysis of RNA sequence structure maps by exhaustive enumeration II. structures of neutral networks and shape space covering. *Monatshefte f. Chemie 127*, 375–389.

Güell, O., F. Sagués, and M. Á. Serrano (2014). Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. *PLoS Comput. Biol. 10*, e1003637.

Haas, J., S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, and T. Schwede (2013). The protein model portala comprehensive resource for protein structure and model information. *Database 2013*, bat031.

Hartl, D. L., A. G. Clark, and A. G. Clark (1997). *Principles of population genetics*, Volume 116. Sinauer associates Sunderland.

Hayden, E. J., E. Ferrada, and A. Wagner (2011). Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature 474*, 92–95.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res. 31*, 3429–3431.

Hofacker, I. L., W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie 125*, 167–188.

Hoque, T., M. Chetty, and A. Sattar (2009). Extended HP model for protein structure prediction. *J. Comput. Biol. 16*, 85–103.

Hosseini, S.-R., A. Barve, and A. Wagner (2015). Exhaustive analysis of a genotype space comprising 10 15 central carbon metabolisms reveals an organization conducive to metabolic innovation. *PLoS Comput. Biol. 11*, e1004329.

Huxley, J. (1942). *Evolution, the modern synthesis.* George Alien and Unwin-, Ltd., London.

Huynen, M. A. (1996). Exploring phenotype space through neutral evolution. *J. Mol. Evol. 43*, 165–169.

Huynen, M. A., D. A. Konings, and P. Hogeweg (1993). Multiple coding and the evolutionary properties of RNA secondary structure. *J. Theor. Biol. 165*, 251–267.

Huynen, M. A., P. F. Stadler, and W. Fontana (1996). Smoothness within ruggedness: The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA 93*, 397–401.

Imamovic, L. and M. O. Sommer (2013). Use of collateral sensitivity networks to design drug cycling protocols that avoid resistance development. *Sci. Transl. Med. 5*, 204ra132.

Innan, H. and F. Kondrashov (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet. 11*, 97–108.

Irbäck, A. and C. Troein (2002). Enumerating designing sequences in the hp model. *J. Biol. Phys. 28*, 1–15.

Jacob, F. (1977). Evolution and tinkering. *Science 196*, 1161–1166.

Jiménez, A., J. Cotterell, A. Munteanu, and J. Sharpe (2015). Dynamics of gene circuits shapes evolvability. *Proc. Natl. Acad. Sci. USA 112*, 2103–2108.

Johannsen, W. (1911). The genotype conception of heredity. *Am. Nat. 45*, 129–159.

Johnston, I. G., S. E. Ahnert, J. P. Doye, and A. A. Louis (2011). Evolutionary dynamics in a simple model of self-assembly. *Phys. Rev. E 83*, 066105.

Jörg, T., O. C. Martin, and A. Wagner (2008). Neutral network sizes of biological RNA molecules can be computed and are not atypically small. *BMC Bioinformatics 9*, 1.

Kauffman, S., C. Peterson, B. Samuelsson, and C. Troein (2003). Random boolean network models and the yeast transcriptional network. *Proc. Natl. Acad. Sci. USA 100*, 14796–14799.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol. 22*, 437–467.

Khersonsky, O. and D. S. Tawfik (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Ann. Rev. Biochem. 79*, 471–505.

Kim, S., T. D. Lieberman, and R. Kishony (2014). Alternating antibiotic treatments constrain evolutionary paths to multidrug resistance. *Proc. Natl. Acad. Sci. USA 111*, 14494–14499.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kivelä, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter (2014). Multilayer networks. *J. Complex Netw. 2*, 203–271.

Koelle, K., S. Cobey, B. Grenfell, and M. Pascual (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science 314*, 1898–1903.

Koonin, E. V. (2011). *The Logic of Chance: The Nature and Origin of Biological Evolution*. FT Press Science. Pearson Education.

Lafforgue, G., F. Martínez, J. Sardanyés, F. de la Iglesia, Q.-W. Niu, S.-S. Lin, R. V. Solé, N.-H. Chua, J.-A. Daròs, and S. F. Elena (2011). Tempo and mode of plant RNA virus escape from RNA interference-mediated resistance. *J. Virol. 85*, 9686–9695.

Lau, K. F. and K. A. Dill (1989). A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules 22*, 3986–3997.

Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler (1991). Long-term experimental evolution in escherichia coli. i. adaptation and divergence during 2,000 generations. *Am. Nat.*, 1315–1341.

Levy, S. B. and B. Marshall (2004). Antibacterial resistance worldwide: causes, challenges and responses. *Nat. Med. 10*, S122–S129.

Li, H., R. Helling, C. Tang, and N. Wingreen (1996). Emergence of preferred structures in a simple model of protein folding. *Science 273*, 666–669.

Lipman, D. J. and W. J. Wilbur (1991). Modelling neutral and selective evolution of protein folding. *Proc. Roy. Soc. London B 245*, 7–11.

Little, S. C., G. Tkačik, T. B. Kneeland, E. F. Wieschaus, and T. Gregor (2011). The formation of the Bicoid morphogen gradient requires protein movement from anteriorly localized mRNA. *PLoS Biol. 9*, e1000596.

Lorenz, R., S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker (2011). ViennaRNA package 2.0. *Algorithms Mol. Biol. 6*, 1.

Manrubia, S. and J. A. Cuesta (2015). Evolution on neutral networks accelerates the ticking rate of the molecular clock. *J. R. Soc. Interface 102*, 20141010.

Manrubia, S. and J. A. Cuesta (2017). Distribution of phenotype sizes in sequence-to-structure genotype-phenotype maps. *arXiv:1702.00351*.

Manrubia, S. C., E. Lázaro, J. Pérez-Mercader, C. Escarmís, and E. Domingo (2003). Fitness distributions in exponentially growing asexual populations. *Phys. Rev. Lett. 90*, 188102.

Maynard Smith, J. (1970). Natural selection and the concept of a protein space. *Nature 225*, 563–564.

Maynard Smith, J., R. Burian, S. Kauffman, P. Alberch, J. Campbell, B. Goodwin, R. Lande, D. Raup, and L. Wolpert (1985). Developmental constraints and evolution: a perspective from the mountain lake conference on development and evolution. *Q. Rev. Biol.*, 265–287.

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press.

Metz, J. A., S. A. Geritz, G. Meszéna, F. J. Jacobs, and J. S. Van Heerwaarden (1995). Adaptive dynamics: a geometrical study of the consequences of nearly faithful reproduction. *IIASA WP 95*, 099.

Milo, R., P. Jorgensen, U. Moran, G. Weber, and M. Springer (2010). Bionumbersthe database of key numbers in molecular and cell biology. *Nucleic Acids Res. 38*, D750–D753.

Montgomery, E., B. Charlesworth, and C. H. Langley (1987). A test for the role of natural selection in the stabilization of transposable element copy number in a population of drosophila melanogaster. *Genet. Res. 49*, 31–41.

Morelli, L. G., K. Uriu, S. Ares, and A. C. Oates (2012). Computational approaches to developmental patterning. *Science 336*, 187–191.

Muñoz-García, J. and S. Ares (2016). Formation and maintenance of nitrogen-fixing cell patterns in filamentous cyanobacteria. *Proc. Natl. Acad. Sci. USA 113*, 6218–6223.

Nei, M. (2013). *Mutation-driven Evolution*. Oxford University Press Oxford.

Niederman, M. S. (2003). Appropriate use of antimicrobial agents: challenges and strategies for improvement. *Crit. Care Med. 31*, 608–616.

O'Brien, P. J. and D. Herschlag (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol. 6*, R91–R105.

Ofria, C. and C. O. Wilke (2004). Avida: A software platform for research in computational evolutionary biology. *Artif. Life 10*, 191–229.

Orth, J. D., T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. Ø. Palsson (2011). A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism2011. *Mol. Sys. Biol. 7*, 535.

Paaby, A. B. and M. V. Rockman (2014). Cryptic genetic variation: evolution's hidden substrate. *Nat. Rev. Genet. 15*, 247–258.

Payne, D. J., M. N. Gwynn, D. J. Holmes, and D. L. Pompliano (2007). Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov. 6*, 29–40.

Payne, J. L., J. H. Moore, and A. Wagner (2014). Robustness, evolvability, and the logic of genetic regulation. *Artif. Life 20*, 111–126.

Payne, J. L. and A. Wagner (2014). Latent phenotypes pervade gene regulatory circuits. *BMC Sys. Biol. 8*, 1.

Peña-Miller, R., D. Laehnemann, G. Jansen, A. Fuentes-Hernandez, P. Rosenstiel, H. Schulenburg, and R. Beardmore (2013). When the most potent combination of antibiotics selects for the greatest bacterial load: the smile-frown transition. *PLoS Biol. 11*, e1001540.

Phillips, R., J. Kondev, J. Theriot, and H. Garcia (2012). *Physical Biology of the Cell*. Garland Science.

Piatigorsky, J. (2007). *Gene Sharing and Evolution: the Diversity of Protein Functions*. Harvard University Press Cambridge MA.

Pigliucci, M. (2008). Sewall wrights adaptive landscapes: 1932 vs. 1988. *Biol. Philos. 23*, 591–603.

Pigliucci, M. and G. B. Müller (2010). *Evolution - the Extended Synthesis*. The MIT Press.

Poelwijk, F. J., D. J. Kiviet, D. M. Weinreich, and S. J. Tans (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature 445*, 383.

Ptashne, M. and A. Gann (2002). *Genes & Signals*, Volume 192. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY.

Radivojac, P., L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, and A. K. Dunker (2007). Intrinsic disorder and functional proteomics. *Biophys. J. 92*, 1439–1456.

Rodrigues, J. F. M. and A. Wagner (2009). Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comp. Biol. 5*, e1000613.

Rodrigues, J. F. M. and A. Wagner (2011). Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Sys. Biol. 5*, 39.

Rodríguez-Rojas, A., M. D. Maciá, A. Couce, C. Gómez, A. Castañeda-García, A. Oliver, and J. Blázquez (2010). Assessing the emergence of resistance: the absence of biological cost in vivo may compromise fosfomycin treatments for p. aeruginosa infections. *PLoS ONE 5*, e10193.

Roemhild, R., C. Barbosa, R. E. Beardmore, G. Jansen, and H. Schulenburg (2015). Temporal variation in antibiotic environments slows down resistance evolution in pathogenic pseudomonas aeruginosa. *Evol. Appl. 8*, 945–955.

Rost, B. et al. (1998). Protein structure prediction in 1d, 2d, and 3d. *Encyclopedia of Computational Chemistry*, 2242–2255.

Rudd, K. E. (2000). EcoGene: a genome sequence database for *Escherichia coli* k-12. *Nucleic Acids Res. 28*, 60–64.

Rué, P. and J. Garcia-Ojalvo (2013). Modeling gene expression in time and space. *Annu. Rev. Biophys. 42*, 605–627.

Salazar-Ciudad, I., J. Jernvall, and S. A. Newman (2003). Mechanisms of pattern formation in development and evolution. *Development 130*, 2027–2037.

Sandler, L. and E. Novitski (1957). Meiotic drive as an evolutionary force. *Am. Nat.*, 105–110.

Schaerli, Y., A. Munteanu, M. Gili, J. Cotterell, J. Sharpe, and M. Isalan (2014). A unified design space of synthetic stripe-forming networks. *Nat. Comm. 5*, 4905.

Schaper, S. and A. A. Louis (2014). The arrival of the frequent: how bias in genotype-phenotype maps can steer populations to local optima. *PLoS ONE 9*, e86635.

Schultes, E. A. and D. P. Bartel (2000). One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science 289*, 448–452.

Schuster, P., W. Fontana, P. F. Stadler, and I. L. Hofacker (1994). From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. London B 255*, 279–284.

Simpson, G. G. (1944). *Tempo and mode in evolution*. Columbia University Press.

Steinberg, B. and M. Ostermeier (2016). Environmental changes bridge evolutionary valleys. *Sci. Adv. 2*, e1500921.

Svensson, E. and R. Calsbeek (2012). *The Adaptive Landscape in Evolutionary Biology*. Oxford University Press.

Tawfik, D. S. (2010). Messy biology and the origins of evolutionary innovations. *Nat. Chem. Biol. 6*, 692.

Tokuriki, N. and D. S. Tawfik (2009). Protein dynamism and evolvability. *Science 324*, 203–207.

Turing, A. M. (1952). The chemical basis of morphogenesis. *Phil. Trans. R. Soc. Lond. B 237*, 37–72.

Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature 150*, 563–565.

Waddington, C. H. (1953). Genetic assimilation of an acquired character. *Evolution*, 118–126.

Wagner, A. (1996). Does evolutionary plasticity evolve? *Evolution*, 1008–1023.

Wagner, A. (2011). *The Origins of Evolutionary Innovations*. Oxford University Press.

Wagner, A. (2014). Mutational robustness accelerates the origin of novel rna phenotypes through phenotypic plasticity. *Biophys. J. 106*, 955–965.

Wagner, A., V. Andriasyan, and A. Barve (2014). The organization of metabolic genotype space facilitates adaptive evolution in nitrogen metabolism. *J. Mol. Biochem. 3*.

Wang, Z. and J. Zhang (2009). Abundant indispensable redundancies in cellular metabolic networks. *Genome Biol. Evol. 1*, 23–33.

Whitehead, D. J., C. O. Wilke, D. Vernazobres, and E. Bornberg-Bauer (2008). The look-ahead effect of phenotypic mutations. *Biol. Direct 3*, 1.

Wolfram, S. (2002). *A New Kind of Science*, Volume 5. Wolfram media Champaign.

Wright, S. (1931). Evolution in mendelian populations. *Genetics 16*, 97–159.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proc. 6th Int. Congr. Genet. 1*, 356–366.

Wuchty, S., W. Fontana, I. L. Hofacker, P. Schuster, et al. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers 49*, 145–165.